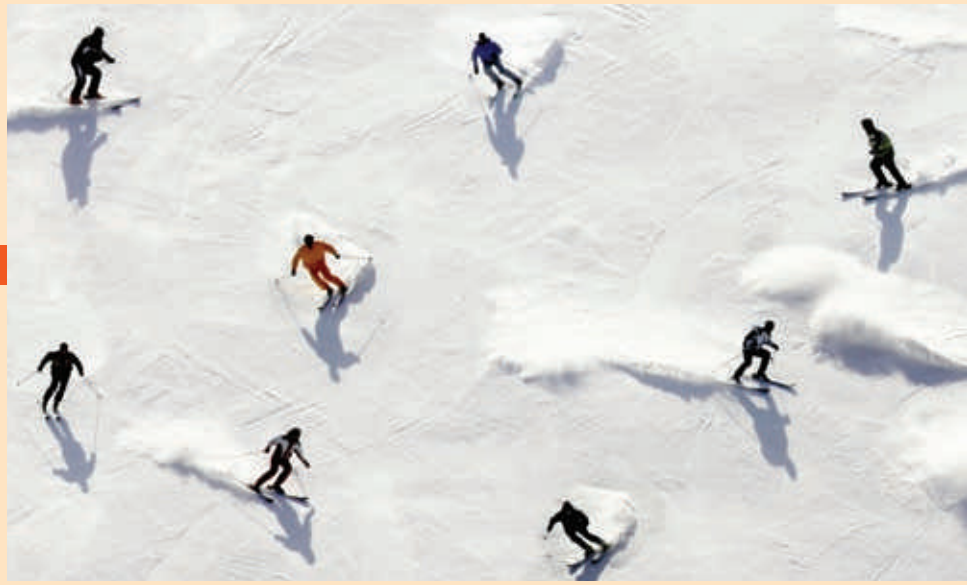
CHAPTER 8

Experiments



CHAPTER OVERVIEW

An experiment is a mode of observation that enables researchers to probe causal relationships. Many experiments in social research are conducted under the controlled conditions of a laboratory, but experimenters can also take advantage of natural occurrences to study the effects of events in the social world.

Introduction

This chapter addresses the controlled experiment: a research method associated more with the natural than the social sciences. We begin Part 3 with this method because the logic and basic techniques of the controlled experiment provide a useful backdrop for understanding other techniques more commonly used in social science, especially for explanatory purposes. We'll also see in this chapter some of the inventive ways social scientists have conducted experiments.

At base, experiments involve (1) taking action and (2) observing the consequences of that action. Social researchers typically select a group of subjects, do something to them, and observe the effect of what was done.

It's worth noting at the outset that we often use experiments in nonscientific inquiry. In preparing a stew, for example, we add salt, taste, add more salt, and taste again. In defusing a bomb, we clip the red wire, observe whether the bomb explodes, clip another, and . . .

We also experiment copiously in our attempts to develop generalized understandings about the world we live in. All skills are learned through experimentation: eating, walking, talking, riding a bicycle, swimming, and so forth. Through experimentation, students discover how much studying is required for academic success. Through experimentation, professors learn how much preparation is required for successful lectures. This chapter discusses how social researchers use experiments to develop generalized understandings. We'll see that, like other methods available to the social researcher, experimenting has its special strengths and weaknesses.

Topics Appropriate for Experiments

Experiments are more appropriate for some topics and research purposes than others. Experiments are especially well suited to research projects involving relatively limited and well-defined concepts and propositions. In terms of the traditional image of science, discussed earlier in this book, the experimental model is especially appropriate for hypothesis testing. Because experiments focus on determining causation, they're also better suited to explanatory than to descriptive purposes.

Let's assume, for example, that we want to discover ways of reducing prejudice against Muslims. We hypothesize that learning about the contribution of Muslims to U.S. history will reduce prejudice, and we decide to test this hypothesis experimentally. To begin, we might test a group of experimental subjects to determine their levels of prejudice against Muslims. Next, we might show them a documentary film depicting the many important ways Muslims have contributed to the scientific, literary, political, and social development of the nation. Finally, we would measure our subjects' levels of prejudice against Muslims to determine whether the film has actually reduced prejudice.

Experimentation has also been successful in the study of small-group interaction. Thus, we might bring together a small group of experimental subjects and assign them a task, such as making recommendations for popularizing car pools. We observe, then, how the group organizes itself and deals with the problem. Over the course of several such experiments, we might systematically vary the nature of the task or the rewards for handling the task successfully. By observing differences in the way groups organize themselves and operate under these varying conditions, we can learn a great deal about the nature of small-group interaction and the factors that influence it. For example, attorneys sometimes present evidence in different ways to different mock juries, to see which method is the most effective.

Political campaigns use experimental methods to determine the most effective types of communication. Different fund-raising messages are evaluated in terms of the funds actually raised.

Laboratory experiments have been used less frequently in the social sciences than in psychology and the natural sciences. Researchers Christine Horne and Michael Lovaglia (2008) argue that this has been a shortcoming in the field of criminology. They have gathered a number of

examples to reveal how laboratory experiments have contributed to understanding with regard to such topics as self-control, social influence, and the law. Horne and Lovaglia do not argue for the replacement of other methods but advocate that studies be augmented with research in laboratory settings.

Similarly, Howard Schuman (2008) details ways in which laboratory experiments can evaluate the effects of differences in question wording and question order in survey research. As we'll see in the next chapter, experienced survey researchers have found differences in public support (or nonsupport) depending on whether government programs are called "welfare" or "assistance to the poor." However, carefully designed experiments can uncover wording impacts that might not be as evident or intuitive to designers of research.

We typically think of experiments as being conducted in laboratories. Indeed, most of the examples in this chapter involve such a setting. This need not be the case, however. Increasingly, social researchers are using the Internet as a vehicle for conducting experiments. Further, sometimes we can construct what are called natural experiments*:* "experiments" that occur in the regular course of social events. The latter portion of this chapter deals with such research.

The Classical Experiment

In both the natural and the social sciences, the most conventional type of experiment involves three major pairs of components: (1) independent and dependent variables, (2) pretesting and posttesting, and (3) experimental and control groups. This section looks at each of these components and the way they're put together in the execution of the experiment.

Independent and Dependent Variables

Essentially, an experiment examines the effect of an independent variable on a dependent variable. Typically, the independent variable takes the form of an experimental stimulus, which is either present or absent. That is, the stimulus is a dichotomous variable, having two attributes, present or not present. In this typical model, the experimenter compares what happens when the stimulus is present to what happens when it is not.

In the example concerning prejudice against Muslims, *prejudice* is the dependent variable and *exposure to Muslim history* is the independent variable. The researcher's hypothesis suggests that prejudice depends, in part, on a lack of knowledge of Muslim history. The purpose of the experiment is to test the validity of this hypothesis by presenting some subjects with an appropriate stimulus, such as a documentary film. In other terms, the independent variable is the cause and the dependent variable is the effect. Thus, we might say that watching the film caused a change in prejudice or that reduced prejudice was an effect of watching the film.

The independent and dependent variables appropriate for experimentation are nearly limitless. Moreover, a given variable might serve as an independent variable in one experiment and as a dependent variable in another. For example, *prejudice* is the dependent variable in our example, but it might be the independent variable in an experiment examining the effect of prejudice on voting behavior.

To be used in an experiment, both independent and dependent variables must be operationally defined. Such operational definitions might involve a variety of observation methods. Responses to a questionnaire, for example, might be the basis for defining prejudice. Speaking to or ignoring Muslims, or agreeing or disagreeing with them, might be elements in the operational definition of interaction with Muslims in a small-group setting.

Conventionally, in the experimental model, dependent and independent variables must be operationally defined before the experiment begins. However, as you'll see in connection with survey research and other methods, it's sometimes appropriate to make a wide variety of observations during data collection and then determine the most useful operational definitions of variables during later analyses. Ultimately, however, experimentation, like other quantitative methods, requires specific and standardized measurements and observations.

Pretesting and Posttesting

In the simplest experimental design, subjects are measured in terms of a dependent variable

(**pretesting**), exposed to a stimulus representing an independent variable, and then remeasured in terms of the dependent variable (**posttesting**). Any differences between the first and last measurements on the dependent variable are then attributed to the independent variable.

In the example of prejudice and exposure to Muslim history, we'd begin by pretesting the extent of prejudice among our experimental subjects. Using a questionnaire asking about attitudes toward Muslims, for example, we could measure both the extent of prejudice exhibited by each individual subject and the average prejudice level of the whole group. After exposing the subjects to the Muslim history film, we could administer the same questionnaire again. Responses given in this posttest would permit us to measure the later extent of prejudice for each subject and the average prejudice level of the group as a whole. If we discovered a lower level of prejudice during the second administration of the questionnaire, we might conclude that the film had indeed reduced prejudice.

In the experimental examination of attitudes such as prejudice, we face a special practical problem relating to validity. As you may already have imagined, the subjects might respond differently to the questionnaires the second time even if their attitudes remain unchanged. During the first administration of the questionnaire, the subjects might be unaware of its purpose. By the second measurement, they might have figured out that the researchers were interested in measuring their prejudice. Because no one wishes to seem prejudiced, the subjects might "clean up" their answers the second time around. Thus, the film would seem to have reduced prejudice although, in fact, it had not.

This is an example of a more general problem that plagues many forms of social research: The very act of studying something may change it. The techniques for dealing with this problem in the context of experimentation will be discussed in various places throughout the chapter. The first technique involves the use of control groups.

Experimental and Control Groups

Laboratory experiments seldom, if ever, involve only the observation of an **experimental group** to which a stimulus has been administered. In addition, the researchers also observe a **control group**, which does not receive the experimental stimulus.

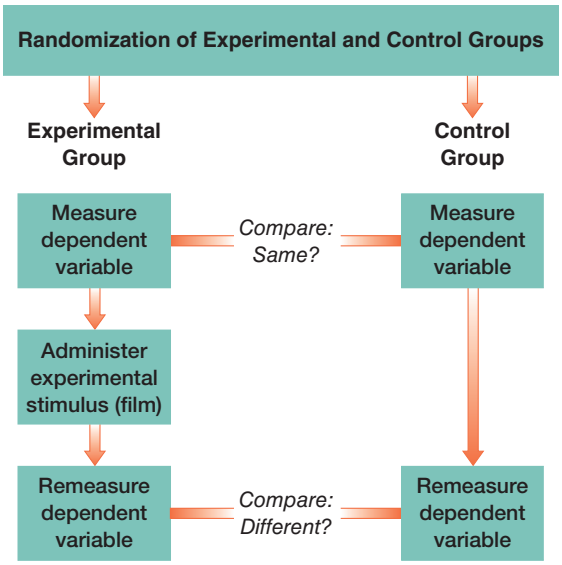


FIGURE 8-1

Diagram of Basic Experimental Design. The fundamental purpose of an experiment is to isolate the possible effect of an independent variable (called the *stimulus* in experiments) on a dependent variable. Members of the experimental group(s) are exposed to the stimulus, whereas those in the control group(s) are not.

In the example of prejudice and Muslim history, we might examine two groups of subjects. To begin, we give each group a questionnaire designed to measure their prejudice against Muslims. Then we show the film to only the experimental group. Finally, we administer a posttest of prejudice to both groups. Figure 8-1 illustrates this basic experimental design.

pretesting The measurement of a dependent variable among subjects.

posttesting The remeasurement of a dependent variable among subjects after they've been exposed to an independent variable.

experimental group In experimentation, a group of subjects to whom an experimental stimulus is administered.

control group In experimentation, a group of subjects to whom no experimental stimulus is administered and who should resemble the experimental group in all other respects. The comparison of the control group and the experimental group at the end of the experiment points to the effect of the experimental stimulus.

Using a control group allows the researcher to detect any effects of the experiment itself. If the posttest shows that the overall level of prejudice exhibited by the control group has dropped as much as that of the experimental group, then the apparent reduction in prejudice must be a function of the experiment or of some external factor rather than a function of the film. If, on the other hand, prejudice is reduced only in the experimental group, this reduction would seem to be a consequence of exposure to the film, because that's the only difference between the two groups. Alternatively, if prejudice is reduced in both groups but to a greater degree in the experimental group than in the control group, that, too, would be grounds for assuming that the film reduced prejudice.

The need for control groups in social research became clear in connection with a series of studies of employee satisfaction conducted by F. J. Roethlisberger and W. J. Dickson (1939) in the late 1920s and early 1930s. These two researchers were interested in discovering what changes in working conditions would improve employee satisfaction and productivity. To pursue this objective, they studied working conditions in the telephone "bank wiring room" of the Western Electric Works in the Chicago suburb of Hawthorne, Illinois.

To the researchers' great satisfaction, they discovered that improving the working conditions increased satisfaction and productivity consistently. As the workroom was brightened up through better lighting, for example, productivity went up. When lighting was further improved, productivity went up again.

To further substantiate their scientific conclusion, the researchers then *dimmed* the lights. Whoops—productivity improved again!

At this point it became evident that the wiring-room workers were responding more to the attention given them by the researchers than to improved working conditions. As a result of this phenomenon, often called the *Hawthorne effect,* social researchers have become more sensitive to and cautious about the possible effects of experiments themselves. In the wiring-room study, the use of a proper control group—one that was studied intensively without any other changes in the working conditions—would have pointed to the presence of this effect.

The need for control groups in experimentation has been nowhere more evident than in medical research. Time and again, patients who participate in medical experiments have appeared to improve, but it has been unclear how much of the improvement has come from the experimental treatment and how much from the experiment. In testing the effects of new drugs, then, medical researchers frequently administer a *placebo*—a "drug" with no relevant effect, such as sugar pills—to a control group. Thus, the control-group patients believe that they, like the experimental group, are receiving an experimental drug. Often, they improve. If the new drug is effective, however, those receiving the actual drug will improve more than those receiving the placebo.

In social science experiments, control groups guard against not only the effects of the experiments themselves but also the effects of any events outside the laboratory during the experiments. In the example of the study of prejudice, suppose that a popular Muslim leader is assassinated in the middle of, say, a weeklong experiment. Such an event may very well horrify the experimental subjects, requiring them to examine their own attitudes toward Muslims, with the result of reduced prejudice. Because such an effect should happen about equally for members of the control and experimental groups, a greater reduction of prejudice among the experimental group would, again, point to the impact of the experimental stimulus: the documentary film.

Sometimes an experimental design requires more than one experimental or control group. In the case of the documentary film, for example, we might also want to examine the impact of reading a book about Muslim history. In that case, we might have one group see the film and read the book, another group only see the movie, still another group only read the book, and the control group do neither. With this kind of design, we could determine the impact of each stimulus separately, as well as their combined effect.

The Double-Blind Experiment

Like patients who improve when they merely think they're receiving a new drug, sometimes experimenters tend to prejudge results. In

medical research, the experimenters may be more likely to "observe" improvements among patients receiving the experimental drug than among those receiving the placebo. (This would be most likely, perhaps, for the researcher who developed the drug.) A **double-blind experiment** eliminates this possibility, because in this design neither the subjects nor the experimenters know which is the experimental group and which is the control. In the medical case, those researchers who were responsible for administering the drug and for noting improvements would not be told which subjects were receiving the drug and which the placebo. Conversely, the researcher who knew which subjects were in which group would not administer the experiment.

In social science experiments, as in medical experiments, the danger of experimenter bias is further reduced to the extent that the operational definitions of the dependent variables are clear and precise. Thus, medical researchers would be less likely to unconsciously bias their reading of a patient's temperature than they would be to bias their assessment of how lethargic the patient was. For the same reason, the small-group researcher would be less likely to misperceive which subject spoke, or to whom he or she spoke, than whether the subject's comments sounded cooperative or competitive, a more subjective judgment that's difficult to define in precise behavioral terms.

The role of the placebo may be more complex than you think, according to a 2010 medical experiment on irritable bowel syndrome. One group of sufferers was given pills in a bottle marked "Placebo" and it was explained that a placebo, sometimes called a sugar pill, contained no active ingredients. Subjects were told that people sometimes seemed to benefit from the placebos. A control group was given no treatment at all. After 21 days the placebo group had improved significantly, while the control group had not.

This study is further complicated, however, by the fact that those receiving the placebo pills also received examinations and counseling sessions, while the control group received no attention at all. Perhaps, as the researchers acknowledge, the positive results were produced by the comprehensive treatment package, not by the placebo pills alone. Also, they note, the measures of improvement were self-assessments. It is possible that physiological measurements might have shown no improvement. But, to complicate matters further, isn't "feeling better" the goal of such treatments?

Selecting Subjects

In Chapter 7 we discussed the logic of sampling, which involves selecting a sample that is representative of some population. Similar considerations apply to experiments. Because most social researchers work in colleges and universities, it seems likely that research laboratory experiments would be conducted with college undergraduates as subjects. Typically, the experimenter asks students enrolled in his or her classes to participate in experiments or advertises for subjects in a college newspaper. Subjects may or may not be paid for participating in such experiments (recall also from Chapter 3 the ethical issues involved in asking students to participate in such studies).

In relation to the norm of generalizability in science, this tendency clearly represents a potential defect in social research. Simply put, college undergraduates are not typical of the public at large. There is a danger, therefore, that we may learn much about the attitudes and actions of college undergraduates but not about social attitudes and actions in general.

However, this potential defect is less significant in explanatory research than in descriptive research. True, having noted the level of prejudice among a group of college undergraduates in our pretesting, we would have little confidence that the same level existed among the public at large. On the other hand, if we found that a documentary film reduced whatever level of prejudice existed among those undergraduates, we would have more confidence—without being certain—that it would have a comparable effect in the community at large. Social processes

double-blind experiment An experimental design in which neither the subjects nor the experimenters know which is the experimental group and which is the control.

and patterns of causal relationships appear to be more generalizable and more stable than specific characteristics such as an individual's level of prejudice.

This problem of generalizing from students isn't always seen as problematic, as Jerome Taylor reports in a commentary on the research into the common cold, a disease he traces back to ancient Egypt. This elusive illness only attacks humans and chimpanzees, so you can probably guess how medical researchers have selected subjects. However, you might be wrong.

> Chimpanzees were too expensive to import en masse, so during the first half of the 20th century British scientists began looking into how the common cold worked by conducting experiments on medical students at St Bartholomew's Hospital in London.
>
> *(Taylor 2008)*

Aside from the question of generalizability, the cardinal rule of subject selection in experimentation concerns the comparability of experimental and control groups. Ideally, the control group represents what the experimental group would be like if it had *not* been exposed to the experimental stimulus. The logic of experiments requires, therefore, that experimental and control groups be as similar as possible. There are several ways to accomplish this.

Probability Sampling

The discussions of the logic and techniques of probability sampling in Chapter 7 provide one method for selecting two groups of people that are similar to each other. Beginning with a sampling frame composed of all the people in the population under study, the researcher might select two probability samples. If these samples each resemble the total population from which they're selected, they'll also resemble each other.

Recall also, however, that the degree of resemblance (representativeness) achieved by probability sampling is largely a function of the sample size. As a general guideline, probability samples of less than 100 are not likely to be terribly representative, and social science experiments seldom involve that many subjects in either experimental or control groups. As a result, then, probability sampling is seldom used in experiments to select subjects from a larger population. Researchers do, however, use the logic of random selection when they assign subjects to groups.

randomization A technique for assigning experimental subjects to experimental and control groups randomly.

Randomization

Having recruited, by whatever means, a total group of subjects, the experimenter may randomly assign those subjects to either the experimental or the control group. The researcher might accomplish such **randomization** by numbering all of the subjects serially and selecting numbers by means of a random number table. Alternatively, the experimenter might assign the odd-numbered subjects to the experimental group and the even-numbered subjects to the control group.

Let's return again to the basic concept of probability sampling. For example, if we use a newspaper advertisement to recruit a total of 40 subjects, there's no reason to believe that these 40 subjects represent the entire population from which they've been drawn. Nor can we assume that the 20 subjects randomly assigned to the experimental group represent that larger population. We can have greater confidence, however, that the 20 subjects randomly assigned to the experimental group will be reasonably similar to the 20 assigned to the control group.

Following the logic of our earlier discussions of sampling, we can see our 40 subjects as a population from which we select two probability samples—each consisting of half the population. Because each sample reflects the characteristics of the total population, the two samples will mirror each other.

As we saw in Chapter 7, our assumption of similarity in the two groups depends in part on the number of subjects involved. In the extreme case, if we recruited only two subjects and assigned, by the flip of a coin, one as the experimental subject and one as the control, there would be no reason to assume that the two subjects are similar to each other. With larger numbers of subjects, however, randomization makes good sense.

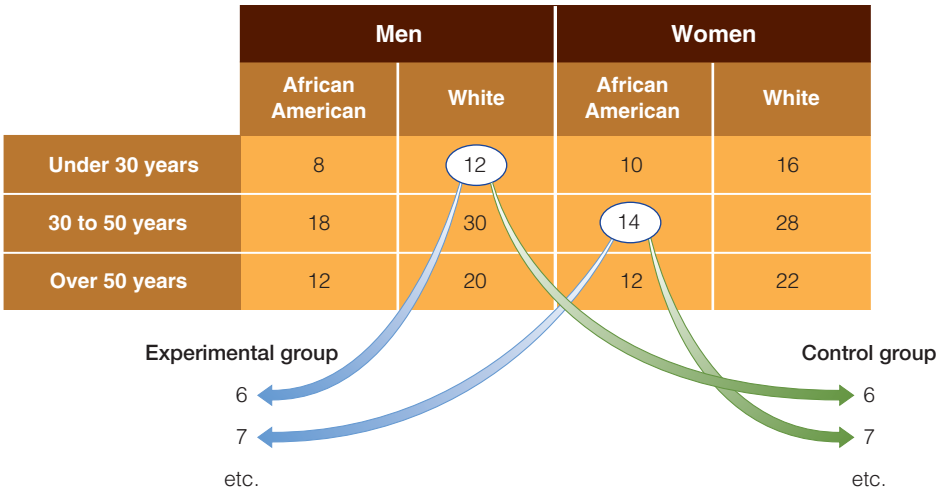


FIGURE 8-2

Quota Matrix Illustration. Sometimes the experimental and control groups are created by finding pairs of matching subjects and assigning one to the experimental group and the other to the control group.

Matching

Another way to achieve comparability between the experimental and control groups is through **matching**. This process is similar to the quota-sampling methods discussed in Chapter 7. If 12 of our subjects are young white men, we might assign 6 of them at random to the experimental group and the other 6 to the control group. If 14 are middle-aged African American women, we might assign 7 to each group. We repeat this process for every relevant grouping of subjects.

The overall matching process could be most efficiently achieved through the creation of a quota matrix constructed of all the most relevant characteristics. Figure 8-2 provides a simplified illustration of such a matrix. In this example, the experimenter has decided that the relevant characteristics are race, age, and gender. Ideally, the quota matrix is constructed to result in an even number of subjects in each cell of the matrix. Then, half the subjects in each cell go into the experimental group and half into the control group.

Alternatively, we might recruit more subjects than our experimental design requires. We might then examine many characteristics of the large initial group of subjects. Whenever we discover a pair of quite similar subjects, we might assign one at random to the experimental group and the other to the control group. Potential subjects who are unlike anyone else in the initial group might be left out of the experiment altogether.

Whatever method we employ, the desired result is the same. The overall average description of the experimental group should be the same as that of the control group. For example, on average both groups should have about the same ages, the same sex composition, the same racial composition, and so forth. This test of comparability should be used whether the two groups are created through probability sampling or through randomization.

Thus far I've referred to the "relevant" variables without saying clearly what those variables are. Of course, these variables cannot be specified in any definite way, any more than I could specify in Chapter 7 which variables should be used in stratified sampling. Which variables are relevant ultimately depends on the nature and purpose of the experiment. As a general rule, however, the control and experimental groups should be comparable in terms of those variables that are most likely to be related to the dependent variable under study. In a study of prejudice, for example, the two groups should be alike in terms of education, ethnicity, and age, among

matching In connection with experiments, the procedure whereby pairs of subjects are matched on the basis of their similarities on one or more variables, and one member of the pair is assigned to the experimental group and the other to the control group.

other characteristics. In some cases, moreover, we may delay assigning subjects to experimental and control groups until we have initially measured the dependent variable. Thus, for example, we might administer a questionnaire measuring subjects' prejudice and then match the experimental and control groups on this variable to assure ourselves that the two groups exhibit the same overall level of prejudice.

Matching or Randomization?

When assigning subjects to the experimental and control groups, you should be aware of two arguments in favor of randomization over matching. First, you may not be in a position to know in advance which variables will be relevant for the matching process. Second, most of the statistics used to analyze the results of experiments assume randomization. Failure to design your experiment that way, then, makes your later use of those statistics less meaningful.

On the other hand, randomization only makes sense if you have a fairly large pool of subjects, so that the laws of probability sampling apply. With only a few subjects, matching would be a better procedure.

Sometimes researchers can combine matching and randomization. When conducting an experiment on the educational enrichment of young adolescents, for example, J. Milton Yinger and his colleagues (1977) needed to assign a large number of students, aged 13 and 14, to several different experimental and control groups to ensure the comparability of students composing each of the groups. They achieved this goal by the following method.

Beginning with a pool of subjects, the researchers first created strata of students nearly identical to one another in terms of some 15 variables. From each of the strata, students were randomly assigned to the different experimental and control groups. In this fashion, the researchers actually improved on conventional randomization. Essentially, they had used a stratified-sampling procedure (Chapter 7), except that they had employed far more stratification variables than are typically used in, say, survey sampling.

Thus far I've described the classical experiment—the experimental design that best represents the logic of causal analysis in the laboratory. In practice, however, social researchers use a great variety of experimental designs. Let's look at some now.

Variations on Experimental Design

Donald Campbell and Julian Stanley (1963), in a classic book on research design, describe 16 different experimental and quasi-experimental designs. This section summarizes a few of these variations to better show the potential for experimentation in social research.

Preexperimental Research Designs

To begin, Campbell and Stanley discuss three "preexperimental" designs, not to recommend them but because they're frequently used in less-than-professional research. These designs are called preexperimental to indicate that they do not meet the scientific standards of experimental designs, and sometimes they may be used because the conditions for full-fledged experiments are impossible to meet. In the first such design—the *one-shot case study*—the researcher measures a single group of subjects on a dependent variable following the administration of some experimental stimulus. Suppose, for example, that we show the Muslim history film, mentioned earlier, to a group of people and then administer a questionnaire that seems to measure prejudice against Muslims. Suppose further that the answers given to the questionnaire seem to represent a low level of prejudice. We might be tempted to conclude that the film reduced prejudice. Lacking a pretest, however, we can't be sure. Perhaps the questionnaire doesn't really represent a sensitive measure of prejudice, or perhaps the group we're studying was low in prejudice to begin with. In either case, the film might have made no difference, though our experimental results might have misled us into thinking it did.

The second preexperimental design discussed by Campbell and Stanley adds a pretest for the experimental group but lacks a control group. This design—which the authors call the *one-group pretest–posttest design*—suffers from the possibility that some factor other than the independent variable might cause a change between

the pretest and posttest results, such as the assassination of a respected Muslim leader. Thus, although we can see that prejudice has been reduced, we can't be sure that the film is what caused that reduction.

To round out the possibilities for preexperimental designs, Campbell and Stanley point out that some research is based on experimental and control groups but has no pretests. They call this design the *static-group comparison*. For example, we might show the Muslim history film to one group and not to another and then measure prejudice in both groups. If the experimental group had less prejudice at the conclusion of the experiment, we might assume the film was responsible. But unless we had randomized our subjects, we would have no way of knowing that the two groups had the same degree of prejudice initially; perhaps the experimental group started out with less.

Figure 8-3 graphically illustrates these three preexperimental research designs by using a different research question: Does exercise cause weight reduction? To make the several designs

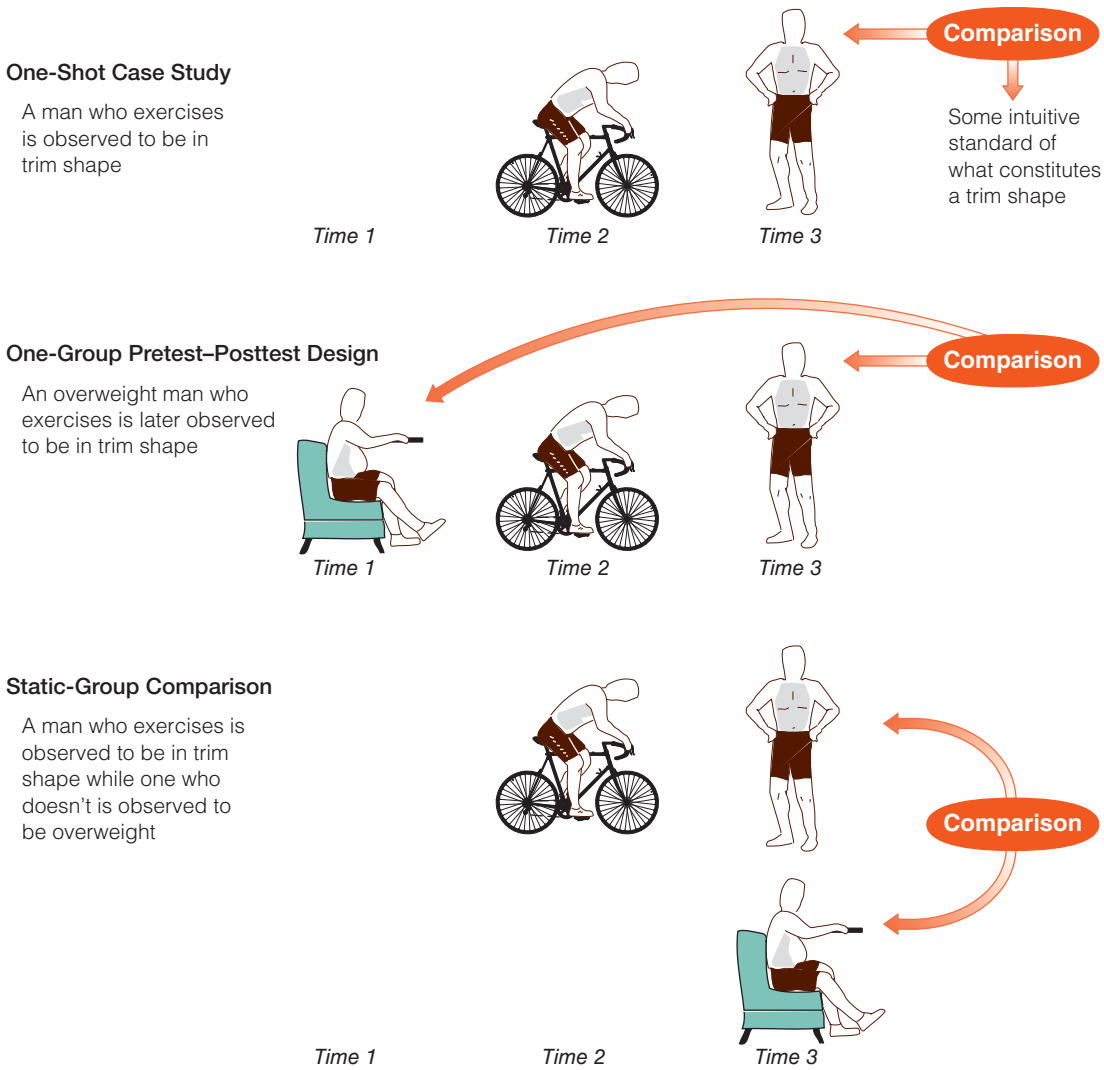


FIGURE 8-3

Three Preexperimental Research Designs. These preexperimental designs anticipate the logic of true experiments but leave themselves open to errors of interpretation. Can you see the errors that might be made in each of these designs? The various risks are solved by the addition of control groups, pretesting, and posttesting.

clearer, the figure shows individuals rather than groups, but the same logic pertains to group comparisons. Let's review the three preexperimental designs in this new example.

The one-shot case study represents a common form of logical reasoning in everyday life. Asked whether exercise causes weight reduction, we may bring to mind an example that would seem to support the proposition: someone who exercises and is thin. There are problems with this reasoning, however. Perhaps the person was thin long before beginning to exercise. Or perhaps he became thin for some other reason, like eating less or getting sick. The observations shown in the diagram do not guard against these other possibilities. Moreover, the observation that the man in the diagram is in trim shape depends on our intuitive idea of what constitutes trim and overweight body shapes. All told, this is very weak evidence for testing the relationship between exercise and weight loss.

The one-group pretest–posttest design offers somewhat better evidence that exercise produces weight loss. Specifically, we've ruled out the possibility that the man was thin before beginning to exercise. However, we still have no assurance that his exercising is what caused him to lose weight.

Finally, the static-group comparison eliminates the problem of our questionable definition of what constitutes trim or overweight body shapes. In this case, we can compare the shapes of the man who exercises and the one who does not. This design, however, reopens the possibility that the man who exercises was thin to begin with. Notice, this is the same as the posttest-only design, mentioned earlier.

Validity Issues in Experimental Research

At this point I want to present, in a more systematic way, the factors that affect the validity of experimental research. First we'll look at what Campbell and Stanley call the sources of internal invalidity, reviewed and expanded in a follow-up book by Thomas Cook and Donald Campbell (1979). Then we'll consider the problem of generalizing experimental results to the "real" world, referred to as external invalidity. Having examined these, we'll be in a position to appreciate the advantages of some of the more sophisticated experimental and quasi-experimental designs social science researchers sometimes use.

internal invalidity Refers to the possibility that the conclusions drawn from experimental results may not accurately reflect what went on in the experiment itself.

Sources of Internal Invalidity

The problem of **internal invalidity** refers to the possibility that the conclusions drawn from experimental results may not accurately reflect what has gone on in the experiment itself. The threat of internal invalidity is present whenever anything other than the experimental stimulus can affect the dependent variable.

Donald Campbell and Julian Stanley (1963: 5–6) and Thomas Cook and Donald Campbell (1979: 51–55) point to several sources of internal invalidity. I will touch on eight of them here to illustrate this concern:

1. *History.* During the course of the experiment, historical events may occur that confound the experimental results. The assassination of a Muslim leader during the course of an experiment on reducing anti–Muslim prejudice is one example.
2. *Maturation.* People are continually growing and changing, and such changes affect the results of the experiment. In a long-term experiment, the fact that the subjects grow older (and wiser?) can have an effect. In shorter experiments, they can grow tired, sleepy, bored, or hungry—or change in other ways that affect their behavior in the experiment.
3. *Testing.* Often the process of testing and retesting influences people's behavior, thereby confounding the experimental results. Suppose we administer a questionnaire to a group as a way of measuring their prejudice. Then we administer an experimental stimulus and remeasure their prejudice. As we saw earlier, by the time we conduct the posttest, the subjects will probably have become more sensitive to the issue of prejudice and will be more thoughtful in their answers. In fact, they may have figured out that we're trying to find out how prejudiced they are, and,

because few people want to appear prejudiced, they may give answers that they think the researchers are seeking or that will make themselves "look good."

4. *Instrumentation.* The process of measurement in pretesting and posttesting brings in some of the issues of conceptualization and operationalization discussed earlier in the book. For example, if we use different measures of the dependent variable (say, different questionnaires about prejudice), how can we be sure they're comparable? Perhaps prejudice will seem to decrease simply because the pretest measure was more sensitive than the posttest measure. Or if the measurements are being made by the experimenters, their standards or abilities may change over the course of the experiment.
5. *Statistical regression.* Sometimes it's appropriate to conduct experiments on subjects who start out with extreme scores on the dependent variable. If you were testing a new method for teaching math to hard-core failures in math, you would want to conduct your experiment on people who previously have done extremely poorly in math. But consider for a minute what's likely to happen to the math achievement of such people over time without any experimental interference. They're starting out so low that they can only stay at the bottom or improve: They can't get worse. Even without any experimental stimulus, then, the group as a whole is likely to show some improvement over time. Referring to a *regression to the mean,* statisticians often point out that extremely tall people as a group are likely to have children shorter than themselves, and extremely short people as a group are likely to have children taller than themselves. There is a danger, then, that changes occurring by virtue of subjects starting out in extreme positions will be attributed erroneously to the effects of the experimental stimulus.
6. *Selection biases.* We discussed selection bias earlier when we examined different ways of selecting subjects for experiments and assigning them to experimental and control groups. Comparisons don't have any meaning unless the groups are comparable at the start of an experiment.
7. *Experimental mortality.* We discussed selection bias earlier when we examined different ways of selecting subjects for experiments and assigning them to experimental and control groups. Comparisons have no meaning unless the groups are comparable at the start of an experiment.
8. *Demoralization.* On the other hand, feelings of deprivation within the control group may result in some giving up. In educational experiments, control-group subjects may feel the experimental group is being treated better and they may become demoralized, stop studying, act up, or get angry.

These, then, are some of the sources of internal invalidity in experiments, as cited by Campbell, Stanley, and Cook. Aware of these pitfalls, experimenters have devised designs aimed at managing them. The classical experiment, coupled with proper subject selection and assignment, addresses each of these problems. Let's look again at that study design, presented in Figure 8-4, as it applies to our hypothetical study of prejudice.

If we use the experimental design shown in Figure 8-4, we should expect two findings from our Muslim history film experiment. For the experimental group, the level of prejudice measured in their posttest should be less than was found in their pretest. In addition, when the two posttests are compared, less prejudice should be found in the experimental group than in the control group.

This design also guards against the problem of history, in that anything occurring outside the experiment that might affect the experimental group should also affect the control group. Consequently, the two posttest results should still differ. The same comparison guards against problems of maturation as long as the subjects have been randomly assigned to the two groups. Testing and instrumentation can't be problems, because both the experimental and control groups are subject to the same tests and experimenter effects. If the subjects have been assigned to the two groups randomly, statistical regression should affect both equally, even if people with extreme scores on prejudice (or whatever the dependent variable is) are being studied. Selection bias is ruled out by the random assignment of subjects. Experimental mortality is more complicated to handle, but the data provided in this

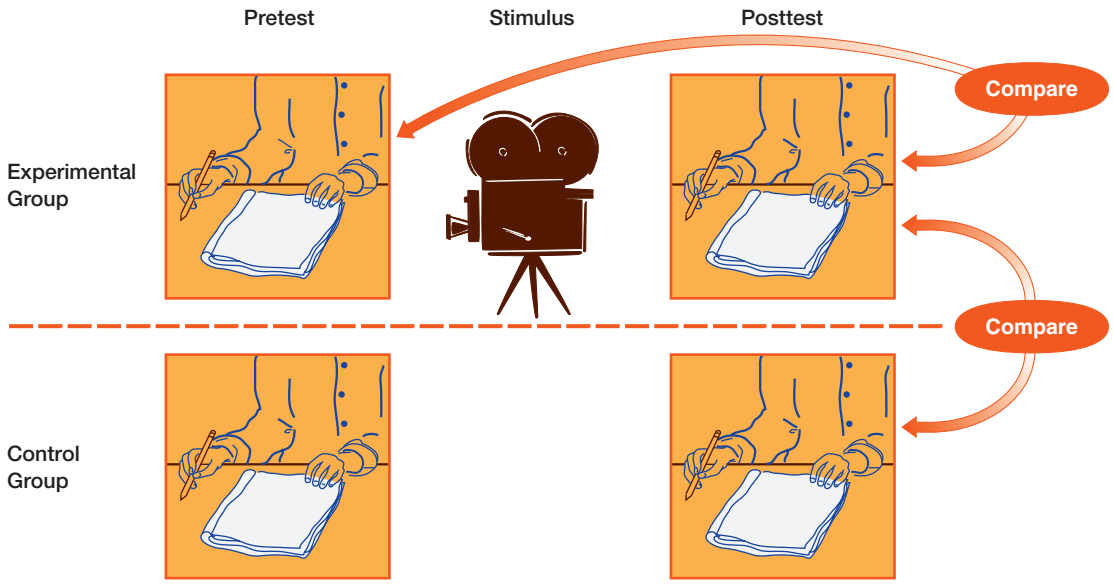


FIGURE 8-4

The Classical Experiment: Using a Muslim History Film to Reduce Prejudice. This diagram illustrates the basic structure of the classical experiment as a vehicle for testing the impact of a film on prejudice. Notice how the control group, the pretesting, and the posttesting function.

study design offer several ways to deal with it. Pretest measurements would let us discover any differences in the dropouts of the experimental and control groups. Slight modifications to the design—administering a placebo (such as a film having nothing to do with Muslims) to the control group, for example—can make the problem even easier to manage. Finally, demoralization can be watched for and taken into account in evaluating the results of the experiment.

Sources of External Invalidity

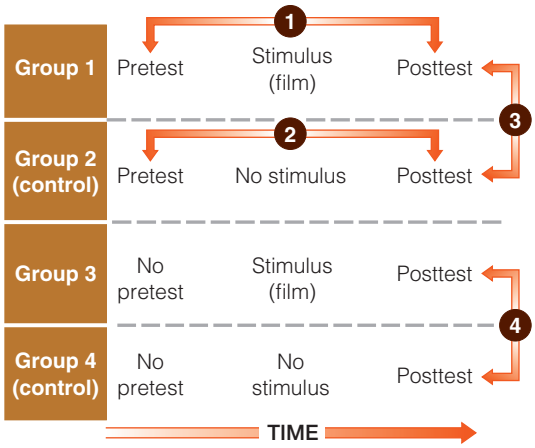
Internal invalidity accounts for only some of the complications faced by experimenters. In addition, there are problems of what Campbell and Stanley call **external invalidity**, which relates to the generalizability of experimental findings to the "real" world. Even if the results of an experiment provide an accurate gauge of what happened during that experiment, do they really tell us anything about life in the wilds of society?

Campbell and Stanley describe four forms of this problem; I'll present one of them as an illustration. The generalizability of experimental findings is jeopardized, as the authors point out, if there's an interaction between the testing situation and the experimental stimulus (1963: 18). Here's an example of what they mean.

Staying with the study of prejudice and the Muslim history film, let's suppose that our experimental group—in the classical experiment—has less prejudice in its posttest than in its pretest and that its posttest shows less prejudice than that of the control group. We can be confident that the film actually reduced prejudice among our experimental subjects. But would it have the same effect if the film were shown in theaters or on television? We can't be sure, because the film might be effective only when people have been sensitized to the issue of prejudice, as the subjects may have been in taking the pretest. This is an example of interaction between the testing and the stimulus. The classical experimental design cannot control for that possibility. Fortunately, experimenters have devised other designs that can.

The *Solomon four-group design* (D. Campbell and Stanley 1963: 24–25) addresses the problem of testing interaction with the stimulus. As the name suggests, it involves four groups of subjects, assigned randomly from a pool. Figure 8-5 presents this design graphically.

external invalidity Refers to the possibility that conclusions drawn from experimental results may not be generalizable to the "real" world.



Expected Findings

1. In Group 1, posttest prejudice should be less than pretest prejudice.
2. In Group 2, prejudice should be the same in the pretest and the posttest.
3. The Group 1 posttest should show less prejudice than the Group 2 posttest does.
4. The Group 3 posttest should show less prejudice than the Group 4 posttest does.

FIGURE 8-5

The Solomon Four-Group Design. The classical experiment runs the risk that pretesting will have an effect on subjects, so the Solomon four-group design adds experimental and control groups that skip the pretest. Thus, it combines the classical experiment and the after-only design (with no pretest).

Notice that Groups 1 and 2 in Figure 8-5 compose the classical experiment, with Group 2 being the control group. Group 3 is administered the experimental stimulus without a pretest, and Group 4 is only posttested. This experimental design permits four meaningful comparisons, which are described in the figure. If the Muslim history film really reduces prejudice—unaccounted for by the problem of internal validity and unaccounted for by an interaction between the testing and the stimulus—we should expect four findings:

1. In Group 1, posttest prejudice should be less than pretest prejudice.
2. In Group 2, prejudice should be the same in the pretest and the posttest.
3. The Group 1 posttest should show less prejudice than the Group 2 posttest.
4. The Group 3 posttest should show less prejudice than the Group 4 posttest.

Notice that Finding 4 rules out any interaction between the testing and the stimulus. And remember that these comparisons are meaningful only if subjects have been assigned randomly to the different groups, thereby providing groups of equal prejudice initially, even though their preexperimental prejudice is measured only in Groups 1 and 2.

There is a side benefit to this research design, as the authors point out. Not only does the Solomon four-group design rule out interactions between testing and the stimulus, it also provides data for comparisons that will reveal how much of this interaction has occurred in a classical experiment. This knowledge allows a researcher to review and evaluate the value of any prior research that used the simpler design.

The last experimental design I'll mention here is what Campbell and Stanley (1963: 25–26) call the *posttest-only control-group design;* it consists of the second half—Groups 3 and 4—of the Solomon design. As the authors argue persuasively, with proper randomization, only Groups 3 and 4 are needed for a true experiment that controls for the problems of internal invalidity as well as for the interaction between testing and stimulus. With randomized assignment to experimental and control groups (which distinguishes this design from the static-group comparison discussed earlier), the subjects will be initially comparable on the dependent variable—comparable enough to satisfy the conventional statistical tests used to evaluate the results—so it's not necessary to measure them. Indeed, Campbell and Stanley suggest that the only justification for pretesting in this situation is tradition. Experimenters have simply grown accustomed to pretesting and feel more secure with research designs that include it. Be clear, however, that this point applies only to experiments in which subjects have been assigned to experimental and control groups randomly, because that's what justifies the assumption that the groups are equivalent without having been measured to find out.

This discussion has introduced the intricacies of experimental design, its problems, and some solutions. There are, of course, a great many other experimental designs in use. Some involve more than one stimulus and combinations of stimuli. Others involve several tests of the dependent variable over time and the administration of the stimulus at different times for different

groups. If you're interested in pursuing this topic, you might want to look at the Campbell and Stanley book.

An Illustration of Experimentation

Experiments have been used to study a wide variety of topics in the social sciences. Some experiments have been conducted within laboratory situations; others occur out in the "real world" and are referred to as *field experiments.* The following discussion provides a glimpse of both. We'll begin with an example of a field experiment.

In George Bernard Shaw's well-loved play *Pygmalion*—the basis of the long-running Broadway musical *My Fair Lady*—Eliza Doolittle speaks of the powers others have in determining our social identity. Here's how she distinguishes the way she's treated by her tutor, Professor Higgins, and by Higgins's friend, Colonel Pickering:

> You see, really and truly, apart from the things anyone can pick up (the dressing and the proper way of speaking, and so on), the difference between a lady and a flower girl is not how she behaves, but how she's treated. I shall always be a flower girl to Professor Higgins, because he always treats me as a flower girl, and always will, but I know I can be a lady to you, because you always treat me as a lady, and always will.
>
> *(Act V)*

The sentiment Eliza expresses here is basic social science, addressed more formally by sociologists such as Charles Horton Cooley (the "looking-glass self") and George Herbert Mead ("the generalized other"). The basic point is that who we think we are—our self-concept—and how we behave are largely a function of how others see and treat us. Related to this, the way others perceive us is largely conditioned by expectations they have in advance. If they've been told we're stupid, for example, they're likely to see us that way—and we may come to see ourselves that way and, in fact, actually act stupidly. "Labeling theory" addresses the phenomenon of people acting in accord with the ways that others perceive and label them. These theories have served as the premise for numerous movies, such as the 1983 film *Trading Places,* in which Eddie Murphy and Dan Aykroyd play a derelict converted into a stockbroker and vice versa.

The tendency to see in others what we've been led to expect takes its name from Shaw's play. Called the "Pygmalion effect," it's nicely suited to controlled experiments. In one of the best-known experimental investigations of the Pygmalion effect, Robert Rosenthal and Lenore Jacobson (1968) administered what they called the "Harvard Test of Inflected Acquisition" to students in a West Coast school. Subsequently, they met with the students' teachers to present the results of the test. In particular, Rosenthal and Jacobson identified certain students as very likely to exhibit a sudden spurt in academic abilities during the coming year, based on the results of the test.

When IQ test scores were compared later, the researchers' predictions proved accurate. The students identified as "spurters" far exceeded their classmates during the following year, suggesting that the predictive test was a powerful one. In fact, the test was a hoax! The researchers had made their predictions randomly among both good and poor students. What they told the teachers did not really reflect students' test scores at all. The progress made by the "spurters" was simply a result of the teachers expecting the improvement and paying more attention to those students, encouraging them, and rewarding them for achievements. (Notice the similarity between this situation and the Hawthorne effect discussed earlier in this chapter.)

The Rosenthal–Jacobson study attracted a great deal of popular as well as scientific attention. Subsequent experiments have focused on specific aspects of what has become known as the *attribution process,* or the *expectations communication model.* This research, largely conducted by psychologists, parallels research primarily by sociologists, which takes a slightly different focus and is often gathered under the label *expectations-states theory*. Psychological studies focus on situations in which the expectations of a dominant individual affect the performance of subordinates—as in the case of a teacher and students, or a boss and employees. The sociological research has tended to focus more on the role of expectations among equals in small,

task-oriented groups. In a jury, for example, how do jurors initially evaluate each other, and how do those initial assessments affect their later interactions? (You can learn more about this phenomenon, including attempts to find practical applications, by searching the web for "Pygmalion effect.")

Here's an example of an experiment conducted to examine the way our perceptions of our abilities and the abilities of others affect our willingness to accept the other person's ideas. Martha Foschi, G. Keith Warriner, and Stephen Hart (1985) were particularly interested in the role "standards" play in that respect:

> In general terms, by "standards" we mean how well or how poorly a person has to perform in order for an ability to be attributed or denied him/her. In our view, standards are a key variable affecting how evaluations are processed and what expectations result. For example, depending on the standards used, the same level of success may be interpreted as a major accomplishment or dismissed as unimportant.
>
> *(1985: 108–9)*

To begin examining the role of standards, the researchers designed an experiment involving four experimental groups and a control. Subjects were told that the experiment involved something called "pattern recognition ability," defined as an innate ability some people had and others did not. The researchers said subjects would be working in pairs on pattern recognition problems.

In fact, of course, there's no such thing as pattern recognition ability. The object of the experiment was to determine how information about this supposed ability affected subjects' subsequent behavior.

The first stage of the experiment was to "test" each subject's pattern recognition abilities. If you had been a subject in the experiment, you would have been shown a geometric pattern for eight seconds, followed by two more patterns, each of which was similar to but not the same as the first one. Your task would be to choose which of the subsequent set had a pattern closest to the first one you saw. You would be asked to do this 20 times, and a computer would print out your "score." Half the subjects would be told that they had gotten 14 correct; the other half would be told that they had gotten only 6 correct—regardless of which patterns they matched with which. Depending on the luck of the draw, you would think you had done either quite well or quite badly. Notice, however, that you wouldn't really have any standard for judging your performance—maybe getting 4 correct would be considered a great performance.

At the same time you were given your score, however, you would also be given your "partner's score," although both the "partners" and their "scores" would also be computerized fictions. (Subjects were told they would be communicating with their partners via computer terminals but would not be allowed to see each other.) If you were assigned a score of 14, you would be told your partner had a score of 6; if you were assigned 6, you would be told your partner had 14.

This procedure meant that you would enter the teamwork phase of the experiment believing either (1) you had done better than your partner or (2) you had done worse than your partner. This information constituted part of the "standard" you would be operating under in the experiment. In addition, half of each group was told that a score of between 12 and 20 meant the subject *definitely* had pattern recognition ability; the other subjects were told that a score of 14 wasn't really high enough to prove anything definite. Thus, you would emerge from this with one of the following beliefs:

1. You are *definitely better* at pattern recognition than your partner.
2. You are *possibly better* than your partner.
3. You are *possibly worse* than your partner.
4. You are *definitely worse* than your partner.

The control group for this experiment was told nothing about their own abilities or those of their partners. In other words, they had no expectations.

The final step in the experiment was to set the "teams" to work. As before, you and your partner would be given an initial pattern, followed by a comparison pair to choose from. When you entered your choice in this round, however, you would be told what your partner had answered; then you would be asked to choose again. In your

final choice, you could either stick with your original choice or switch. The "partner's" choice was, of course, created by the computer, and as you can guess, there were often disagreements in the teams: 16 out of 20 times, in fact.

The dependent variable in this experiment was the extent to which subjects would switch their choices to match those of their partners. The researchers hypothesized that the *definitely better* group would switch least often, followed by the *possibly better* group, followed by the *control group,* followed by the *possibly worse* group, followed by the *definitely worse* group, who would switch most often.

The number of times subjects in the five groups switched their answers follows. Realize that each had 16 opportunities to do so. These data indicate that each of the researchers' expectations was correct—with the exception of the comparison between the *possibly worse* and *definitely worse* groups. Although the latter group was in fact the more likely to switch, the difference was too small to be taken as a confirmation of the hypothesis. (Chapter 16 will discuss the statistical tests that let researchers make decisions like this.)

Group	*Mean Number of Switches*
Definitely better	5.05
Possibly better	6.23
Control group	7.95
Possibly worse	9.23
Definitely worse	9.28

In more-detailed analyses, it was found that the same basic pattern held for both men and women, though it was somewhat clearer for women than for men. Here are the actual data:

	Mean Number of Switches	
	Women	*Men*
Definitely better	4.50	5.66
Possibly better	6.34	6.10
Control group	7.68	8.34
Possibly worse	9.36	9.09
Definitely worse	10.00	8.70

Because specific research efforts like this one sometimes seem extremely focused in their scope, you might wonder about their relevance to anything. As part of a larger research effort, however, studies like this one add concrete pieces to our understanding of more-general social processes.

It's worth taking a minute to consider some of the life situations where "expectation states" might have very real and important consequences. I've mentioned the case of jury deliberations. How about all forms of prejudice and discrimination? Or, consider how expectation states figure into job interviews or meeting your heartthrob's parents. If you think about it, you'll undoubtedly see other situations where these laboratory concepts apply in real life.

Alternative Experimental Settings

Although we tend to equate the terms *experiment* and *laboratory experiment,* many important social science experiments occur outside controlled settings, as we've seen in our example of the Rosenthal–Jacobson study of the Pygmalion effect. Two other special circumstances deserve mention here: web-based experiments and "natural" experiments.

Here's a different kind of social science experiment. Shelley J. Correll, Stephen Benard, and In Paik (2007) were interested in learning whether race, gender, and/or parenthood might produce discrimination in hiring. Specifically, they wanted to find out if there was a "Motherhood penalty." These researchers decided to explore this topic with an experiment using college undergraduates. The student-subjects chosen for the study were told that a new communications company was looking for someone to manage the marketing department of their East Coast office.

> They heard that the communications company was interested in receiving feedback from younger adults since young people are heavy consumers of communications technology. To further increase their task orientation, participants were told that their input would be incorporated with the other information the company collects on applicants and would impact actual hiring decisions.
>
> *(2007: 1311)*

The researchers had created a number of resumes describing fictitious candidates for the manager's position. Initially, the resumes had no indication of race, sex, or parenthood, and a group of subjects was asked to evaluate the quality of the candidates. The initial evaluations showed the resumes to be equivalent in apparent quality.

Then, in the main experiment, the resumes were augmented with additional information. Gender became apparent when names were added to the resumes. Moreover, the use of typically African American names (e.g., Latoya and Ebony for women; Tyrone and Jamal for men) or typically white names (e.g., Allison and Sarah for women; Brad and Matthew for men) allowed subjects to guess the candidates' races. Finally, listing participation in a Parent–Teacher Association or listing names of children identified some candidates as parents. Over the course of the experiment, these different status indicators were added to the same resumes. Thus a particular resume might appear as a black mother, a white non-mother, a white father, and so forth. Of course, no student-subject would evaluate the same resume with different status indicators.

Finally, the experimental subjects were given sets of resumes to evaluate in a number of ways. For example, they were asked how competent they felt the candidates were and how committed they seemed. They were asked to suggest a salary that might be offered a given candidate and to predict how likely it was that the candidate would eventually be promoted within the organization. They were even asked to indicate how many days the candidate should be allowed to miss work or come late before being fired.

Since each of the resumes was evaluated with different status indicators attached, it was possible for the experimenters to determine whether those statuses made a difference. Specifically, they could test for the existence of a Motherhood penalty. And they found it. Among other things:

- Mothers were judged less competent and less committed than non-mothers.
- Students offered the mothers lower salaries than the non-mothers and would allow them fewer missed or late days on the job.
- They felt the mothers were less likely to be promoted than the non-mothers.
- And they were almost twice as likely to recommend hiring the non-mothers.

Rounding out the analysis of gender and parenthood, the researchers found that, while the differences were smaller for men than for women, fathers were rated *higher* than non-fathers. This was just the opposite pattern as had been found among women candidates.

The Motherhood penalty was found among both white and African American candidates. Moreover, it did not matter what the gender of the subject evaluators were. Both women and men rated mothers lower than non-mothers.

Factorial Designs

Up to now, I have discussed the experimental variable as singular: We try to limit the variation between experimental and control group to one variable. While this logic is basic to the experimental model, **factorial designs** expand that model to encompass more than one experimental variable. Let's say we are interested in what brings consumers to hunger for Green Healthy Treats (GHT). Are they more moved by environmental or health issues?

Let's suppose we create TV spots that (1) emphasize the environmental value of the way GHT is produced and (2) and how healthy it is for you. We produce two ads, let's call them E and H to reflect Environmental and Health emphases. Now, instead of having one experimental group, we have three:

E only

H only

E & H both

Now we can compare the desire for GHT among those who were shown the Environmental ad only (E), the Health ad only (H), and both ads (E & H). This design enables us to determine whether (a) the Environmental ad makes a difference, regardless of whether viewers saw the Health ad; (b) the Environmental ad makes a difference regardless of whether they saw the Environmental ad; (c) these two ads have independent, cumulative support for using GHT; or (d) neither ad makes a difference.

factorial design An experimental design using more than one experimental variable.

Web-Based Experiments

Increasingly, researchers are using the Internet as a vehicle for conducting social science experiments. Because representative samples are not essential in most experiments, researchers can often use volunteers who respond to invitations online. One site you might visit to get a better idea of this form of experimentation is Online Social Psychology Studies. This website offers hot links to numerous professional and student research projects on such topics as "interpersonal relations," "beliefs and attitudes," and "personality and individual differences." In addition, the site offers some resources for conducting web experiments.

"Natural" Experiments

Important social science experiments can occur in the course of normal social events, outside controlled settings. Sometimes nature designs and executes experiments that we can observe and analyze; sometimes social and political decision makers serve this natural function.

Imagine, for example, that a hurricane has struck a particular town. Some residents of the town suffer severe financial damages, and others escape relatively lightly. What, we might ask, are the behavioral consequences of suffering a natural disaster? Are those who suffer most more likely to take precautions against future disasters than are those who suffer least? To answer these questions, we might interview residents of the town some time after the hurricane. We might question them regarding the precautions they had taken before the hurricane and those they're currently taking toward future preparedness. We could then compare the precautionary actions of the people who suffered a great deal from the hurricane with those taken by citizens who suffered relatively little. In this fashion, we might take advantage of a natural experiment, which we could not have arranged even if we'd been perversely willing to do so.

Because the researcher must, for the most part, take things as they occur, natural experiments raise many of the validity problems discussed earlier. Thus, when Stanislav Kasl, Rupert Chisolm, and Brenda Eskenazi (1981) chose to study the impact that the Three Mile Island (TMI) nuclear accident in Pennsylvania had on plant workers, they had to be especially careful while devising the study design:

> Disaster research is necessarily opportunistic, quasi-experimental, and after-the-fact. In the terminology of Campbell and Stanley's classical analysis of research designs, our study falls into the "static-group comparison" category, considered one of the weak research designs. However, the weaknesses are potential and their actual presence depends on the unique circumstances of each study.
>
> *(1981: 474)*

The foundation of this study was a survey of the people who had been working at Three Mile Island on March 28, 1979, when the cooling system failed in the number 2 reactor and began melting the uranium core. The survey was conducted five to six months after the accident. Among other things, the survey questionnaire measured workers' attitudes toward working at nuclear power plants. If they had measured only the TMI workers' attitudes after the accident, the researchers would have had no idea whether attitudes had changed as a consequence of the accident. But they improved their study design by selecting another, nearby—seemingly comparable—nuclear power plant (abbreviated as PB) and surveyed workers there as a control group: hence their reference to a static-group comparison.

Even with an experimental and a control group, the authors were wary of potential problems in their design. In particular, their design was based on the idea that the two sets of workers were equivalent to each other, except for the single fact of the accident. The researchers could have assumed this if they had been able to assign workers to the two plants randomly, but of course that was not the case. Instead, they needed to compare characteristics of the two groups and infer whether or not they were equivalent. Ultimately, the researchers concluded that the two sets of workers were very much alike, and the plant the employees worked at was merely a function of where they lived.

Even granting that the two sets of workers were equivalent, the researchers faced another problem of comparability. They could not contact all the workers who had been employed at TMI at the time of the accident. The researchers discussed the problem as follows:

> One special attrition problem in this study was the possibility that some of the no-contact nonrespondents among the TMI subjects, but not PB subjects, had permanently left the area because of the accident. This biased attrition would, most likely, attenuate the estimated extent of the impact. Using the evidence of disconnected or "not in service" telephone numbers, we estimate this bias to be negligible (1 percent).
>
> *(Kasl, Chisolm, and Eskenazi 1981: 475)*

The TMI example points to both the special problems involved in natural experiments and the possibility for taking those problems into account. Social research generally requires ingenuity and insight, and natural experiments are certainly no exception. Earlier in this chapter, we used a hypothetical example of studying whether an ethnic history film reduced prejudice. Sandra Ball-Rokeach, Joel Grube, and Milton Rokeach (1981) were able to address that topic in real life through a natural experiment. In 1977, the television dramatization of Alex Haley's *Roots*, a historical saga about African Americans, was presented by ABC on eight consecutive nights. It garnered the largest audiences in television history up to that time. Ball-Rokeach and her colleagues wanted to know whether *Roots* changed white Americans' attitudes toward African Americans. Their opportunity arose in 1979, when a sequel—*Roots: The Next Generation*—was televised. Although it would have been nice (from a researcher's point of view) to assign random samples of Americans either to watch or not to watch the show, that wasn't possible. Instead, the researchers selected four samples in Washington State and mailed questionnaires that measured attitudes toward African Americans. Following the last episode of the show, respondents were called and asked how many, if any, episodes they had watched. Subsequently, questionnaires were sent to respondents, remeasuring their attitudes toward African Americans.

By comparing attitudes before and after for both those who watched the show and those who didn't, the researchers reached several conclusions. For example, they found that people with already egalitarian attitudes were much more likely to watch the show than were those who were more prejudiced toward African Americans: a self-selection phenomenon. Comparing the before and after attitudes of those who watched the show, moreover, suggested the show itself had little or no effect. Those who watched it were no more egalitarian afterward than they had been before.

This example anticipates the subject of Chapter 12, evaluation research, which can be seen as a special type of natural experiment. As you'll see, evaluation research involves taking the logic of experimentation into the field to observe and evaluate the effects of stimuli in real life. Because this is an increasingly important form of social research, an entire chapter is devoted to it.

Strengths and Weaknesses of the Experimental Method

Experiments are the primary tool for studying causal relationships. However, like all research methods, experiments have both strengths and weaknesses.

The chief advantage of a controlled experiment lies in the isolation of the experimental variable's impact over time. This is seen most clearly in terms of the basic experimental model. A group of experimental subjects are found, at the outset of the experiment, to have a certain characteristic; following the administration of an experimental stimulus, they are found to have a different characteristic. To the extent that subjects have experienced no other stimuli, we may conclude that the change of characteristics is attributable to the experimental stimulus.

Further, because individual experiments are often rather limited in scope, requiring relatively little time and money and relatively few subjects, we often can replicate a given experiment several times using several different groups of subjects. (This isn't always the case, of course, but it's usually easier to repeat experiments than, say, surveys.) As in all other forms of scientific research, replication of research findings strengthens our confidence in the validity and generalizability of those findings.

The greatest weakness of laboratory experiments lies in their artificiality. Social processes that occur in a laboratory setting might not necessarily occur in natural social settings. For example, a Muslim history film might genuinely reduce prejudice among a group of experimental

subjects. This would not necessarily mean, however, that the same film shown in neighborhood movie theaters throughout the country would reduce prejudice among the general public. Artificiality is not as much of a problem, of course, for natural experiments as for those conducted in the laboratory.

In discussing several of the sources of internal and external invalidity mentioned by Campbell, Stanley, and Cook, we saw that we can create experimental designs that logically control such problems. This possibility points to one of the great advantages of experiments: They lend themselves to a logical rigor that is often much more difficult to achieve in other modes of observation.

Ethics and Experiments

As you've probably realized by now, researchers must consider many important ethical issues in conducting social science experiments. I'll mention only two here.

First, experiments almost always involve deception. In most cases, explaining the purpose of the experiment to subjects would probably cause them to behave differently—trying to look less prejudiced, for example. It's important, therefore, to determine (1) whether a particular deception is essential to the experiment and (2) whether the value of what may be learned from the experiment justifies the ethical violation.

Second, experiments are typically intrusive. Subjects often are placed in unusual situations and asked to undergo unusual experiences. Even when the subjects are not physically injured (don't do that, by the way), there is always the possibility that they could be psychologically damaged, as some of the previous examples in this chapter have illustrated. As with the matter of deception, you'll find yourself balancing the potential value of the research against the potential damage to subjects.

MAIN POINTS

Introduction

- In experiments, social researchers typically select a group of subjects, do something to them, and observe the effect of what was done.

Topics Appropriate for Experiments

- Experiments are an excellent vehicle for the controlled testing of causal processes.

The Classical Experiment

- The classical experiment tests the effect of an experimental stimulus (the independent variable) on a dependent variable through the pretesting and posttesting of experimental and control groups.
- It is generally less important that a group of experimental subjects be representative of some larger population than that experimental and control groups be similar to each other.
- A double-blind experiment guards against experimenter bias, because neither the experimenter nor the subject knows which subjects are in the control group(s) and which are in the experimental group(s).

Selecting Subjects

- Probability sampling, randomization, and matching are all methods of achieving comparability in the experimental and control groups. Randomization is the generally preferred method. In some designs, it can be combined with matching.

Variations on Experimental Design

- Campbell and Stanley describe three forms of preexperiments: the one-shot case study, the one-group pretest–posttest design, and the static-group comparison. None of these designs features all the controls available in a true experiment.
- Campbell and Stanley list, among others, eight sources of internal invalidity in experimental design. The classical experiment with random assignment of subjects guards against each of these problems.
- Experiments also face problems of external invalidity: Experimental findings may not reflect real life.
- The interaction of testing and stimulus is an example of external invalidity that the classical experiment does not guard against.
- The Solomon four-group design and other variations on the classical experiment can safeguard against external invalidity.
- Campbell and Stanley suggest that, given proper randomization in the assignment of subjects to the experimental and control groups, there is no need for pretesting in experiments.

An Illustration of Experimentation

- Experiments on "expectation states" demonstrate experimental designs and show how experiments can prove relevant to real-world concerns.

Alternative Experimental Settings

- More and more, researchers are using the Internet for conducting experiments.
- Natural experiments often occur in the course of social life in the real world, and social researchers can implement them in somewhat the same way they would design and conduct laboratory experiments.

Strengths and Weaknesses of the Experimental Method

- Like all research methods, experiments have strengths and weaknesses. Their primary weakness is artificiality: What happens in an experiment may not reflect what happens in the outside world. Strengths include the isolation of the independent variable, which permits causal inferences; the relative ease of replication; and scientific rigor.

Ethics and Experiments

- Experiments typically involve deceiving subjects.
- By their intrusive nature, experiments open the possibility of inadvertently causing damage to subjects.

KEY TERMS

The following terms are defined in context in the chapter and at the bottom of the page where the term is introduced, as well as in the comprehensive glossary at the back of the book.

control group
double-blind experiment
experimental group
external invalidity
factorial design
internal invalidity
matching
posttesting
pretesting
randomization

PROPOSING SOCIAL RESEARCH: EXPERIMENTS

In the next series of exercises, we'll focus on specific data-collection techniques, beginning with experiments here. If you're doing these exercises as part of an assignment in the course, your instructor will tell you whether you should skip those chapters dealing with methods you won't use. If you're doing these exercises on your own, to improve your understanding of the topics in the book, you can temporarily modify your proposed data-collection method and explore how you would research your topic using the method at hand—in this case, experimentation.

In the proposal, you'll describe the experimental stimulus and how it will be administered, as well as detailing the experimental and control groups you'll use. You'll also describe the pretesting and posttesting that will be involved in your experiment. What will be the setting for your experiments: a laboratory or more-natural circumstances?

It may be appropriate for you to conduct a double-blind experiment, in which case you should describe how you will accomplish it. You may also need to explore some of the internal and external problems of validity that might complicate your analysis of your results.

Finally, the experimental model is used to test specific hypotheses, so you should detail how you will accomplish that in terms of your study.

REVIEW QUESTIONS AND EXERCISES

1. In the library or on the web, locate a research report of an experiment. Identify the dependent variable and the stimulus.
2. Pick 4 of the 8 sources of internal invalidity discussed in this chapter and make up examples (not discussed in the chapter) to illustrate each.
3. Create a hypothetical experimental design that illustrates one of the problems of external invalidity.
4. Think of a recent natural disaster you've witnessed or read about. Frame a research question that might be studied by treating that disaster as a natural experiment. In two or three paragraphs, outline how the study might be done.
5. In this chapter, we looked briefly at the problem of "placebo effects." On the web, find a study in which the placebo effect figured importantly. Write a brief report on the study, including the source of your information. (*Hint:* You might want to do a search on "placebo.")