CHAPTER 8

Experimental Research

INTRODUCTION

Experimental research builds on the principles of a positivist approach more directly than do the other research techniques. Researchers in the natural sciences (e.g., chemistry and physics), related applied fields (e.g., agriculture, engineering, and medicine), and the social sciences conduct experiments. The logic that guides an experiment on plant growth in biology or testing a metal in engineering is applied in experiments on human social behavior. Although it is most widely used in psychology, the experiment is found in education, criminal justice, journalism, marketing, nursing, political science, social work, and sociology. This chapter focuses first on the experiment conducted in a laboratory under controlled conditions, then looks at experiments conducted in the field.

The experiment's basic logic extends commonsense thinking. Commonsense experiments are less careful or systematic than scientifically based experiments. In commonsense language, an *experiment* means modifying something in a situation, then comparing an outcome to what existed without the modification. For example, I try to start my car. To my surprise, it does not start. I "experiment" by cleaning off the battery connections, then try to start it again. I modified something (cleaned the connections) and compared the outcome (whether the car started) to the previous situation (it did not start). I began with an implicit "hypothesis"—a buildup of crud on the connections is the reason the car is not starting, and once the crud is cleaned off, the car will start. This illustrates three things researchers do in experiments: (1) begin with a hypothesis, (2) modify something in a situation, and (3) compare outcomes with and without the modification.

Compared to the other social research techniques, experimental research is the strongest for testing causal relationships because the three conditions for causality (temporal order, association, and no alternative explanations) are clearly met in experimental designs.

Research Questions Appropriate for an Experiment

The Issue of an Appropriate Technique. Social researchers use different research techniques (e.g., experiments and surveys) because some research questions can be addressed with certain techniques but not with others. New researchers often ask which research technique best fits which problem. This is difficult to answer because there is no fixed match between problem and technique. The answer is: Make an informed judgment.

General guidelines exist for fitting techniques to problems. Beyond guidelines, you can develop judgment from reading research reports, understanding the strengths and weaknesses of different techniques, assisting more experienced researchers with their research, and gaining practical experience.

Research Questions for Experimental Research. The experiment is a powerful way to focus sharply on causal relations, and it has practical advantages over other techniques, but it also has limitations. The research questions most appropriate for an experiment fit its strengths and limitations. These include its basic logic and practical restrains, its narrow scope, its ability to isolate causes, and the convention of researchers.

The questions appropriate for using an experimental logic confront ethical and practical limitations of intervening in human affairs for research purposes. It is immoral or impossible to manipulate many areas of human life for research purposes. The pure logic of an experiment has an experimenter intervene or induce a change in some focused part of social life, then examine the consequences that result from the change or intervention. This usually means that the experiment is limited to research questions in which a researcher is able to manipulate conditions. Experimental research cannot answer questions such as, Do people who complete a college education increase their annual income

more than people who do not? Do children raised with younger siblings develop better leadership skills than only children? Do people who belong to more organizations vote more often in elections? This is because an experimenter often cannot manipulate conditions or intervene. He or she cannot randomly assign thousands to attend college and prevent others from attending to discover who later earns more income. He or she cannot induce couples to have either many children or a single child so he or she can examine how leadership skills develop in children. He or she cannot compel people to join or quit organizations then see whether they vote. Experimenters are highly creative in simulating such interventions or conditions, but they cannot manipulate many of the variables of interest to fit the pure experimental logic.

The experiment is usually best for issues that have a narrow scope or scale. This strength allows experimenters to assemble and "run" many experiments with limited resources in a short period. Some carefully designed experiments require assembling only 50 or 60 volunteers and can be completed in one or two months. In general, the experiment is better suited for micro-level (e.g., individual or small-group phenomena) than for macro-level theoretical concerns or questions. This is one reason why psychologists, social psychologists in sociology, and political psychologists in political science all tend to use experiments. Experiments can rarely address questions that require looking at conditions across an entire society or across decades. Use of the experiment may limit the types of variables that one can examine, the questions that one can address, and one's ability to generalize to larger settings (see External Validity and Field Experiments later in this chapter).

Experiments encourage researchers to isolate and target the impact that arises from one or a few causal variables. This strength in demonstrating causal effects is a limitation in situations where a researcher tries to examine numerous variables simultaneously. The experiment is rarely appropriate for research questions or issues that require a researcher to examine the impact of dozens of diverse variables all together. Rarely do experiments permit assessing conditions across a wide range of complex settings or numerous social groups all at the same time. Although the accumulated knowledge from many individual experiments, each focused on one or two variables, may advance understanding, the expermiment is different from research on a highly complex situation that tries to examine how dozens of variables operate simultaneously.

A last factor that influences the research questions that fit the experimental method is convention. For some topics or research questions, numerous researchers depended on the experimental method to create a large body of literature with hundreds of studies. This facilitates quick, smooth communication. More importantly, it allows researchers to advance knowledge rapidly by replicating previous experiments with only minor adjustments in study design and to isolate precisely the effects of specific conditions or variables. It is a limitation because those who specialize in a topic will tend to evaluate all new research by the criteria of a good experiment. However, it does not mean that a study using a technique different from the experiment is inappropriate; out of habit or convention specialists in the area will be more critical of it and slower to accept and assimilate new knowledge from a nonexperimental study.

Often, it is possible to conduct research on closely related topics using either an experimental or a nonexperimental method. For example, a researcher may wish to study attitudes toward people in wheelchairs. An experimenter might ask people to respond (e.g., Would you hire this person? How comfortable would you be if this person asked you for a date?) to photos of some people in wheelchairs and some people not in wheelchairs. A survey researcher might ask people their opinions about people in wheelchairs. The field researcher might observe people's reactions to someone in a wheelchair, or the researcher himself or herself might be in wheelchair and carefully note the reactions of others.

RANDOM ASSIGNMENT

Social researchers frequently want to compare. For example, a researcher has two groups of 15 students and wants to compare the groups on the basis of a key difference between them (e.g., a course that one group completed). Or a researcher has five groups of customers and wants to compare the groups on the basis of one characteristic (e.g., geographic location). The cliché, "Compare apples to apples, don't compare apples to oranges," is not about fruit; it is about comparisons. It means that a valid comparison depends on comparing things that are fundamentally alike. Random assignment facilitates comparison in experiments by creating similar groups.

When making comparisons, researchers want to compare cases that do not differ with regard to variables that offer alternative explanations. For example, a researcher compares two groups of students to determine the impact of completing a course. In order to be compared, the two groups must be similar in most respects except for taking the course. If the group that completed the course is also older than the group that did not, for example, the researcher cannot determine whether completing the course or being older accounts for differences between the groups.

Why Randomly Assign?

Random assignment is a method for assigning cases (e.g., individuals, organizations, etc.) to groups for the purpose of making comparisons. It is a way to divide or sort a collection of cases into two or more groups in order to increase one's confidence that the groups do not differ in a systematic way. It is a mechanical method; the assignment is automatic, and the researcher cannot make assignments on the basis of personal preference or the features of specific cases.

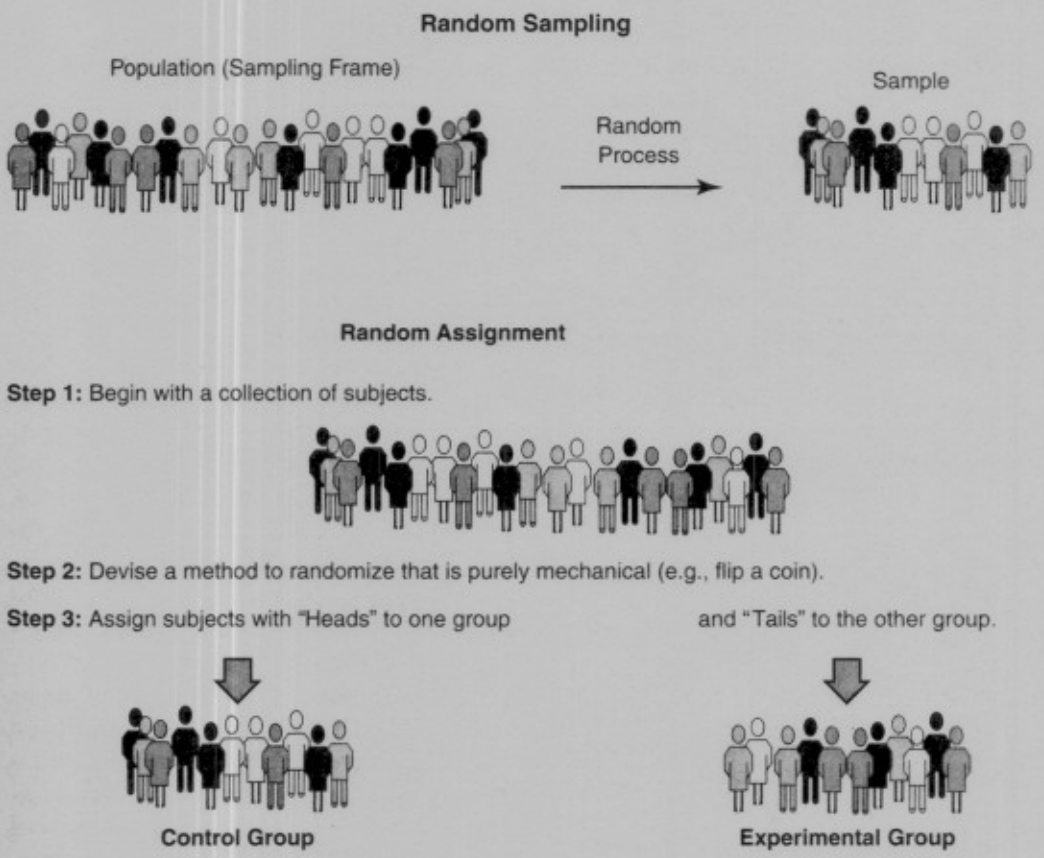
Random assignment is random in a statistical or mathematical sense, not in an everyday sense. In everyday speech, *random* means unplanned, haphazard, or accidental, but it has a specialized meaning in mathematics. In probability theory, *random* describes a process in which each case has a known chance of being selected. Random selection lets a researcher calculate the odds that a specific case will be sorted into one group over another. Thus, the selection process obeys mathematical laws, which makes precise calculations possible. For example, a random process is one in which all cases have an exactly equal chance of ending up in one or the other group.

The wonderful thing about a random process is that over many separate random occurrences, predictable things happen. Although the process is entirely due to chance and it is impossible to predict a specific outcome at a specific time, very accurate predictions are possible over many situations.

Random assignment or randomization is unbiased because a researcher's desire to confirm a hypothesis or a research subject's personal interests do not enter into the selection process. *Unbiased* does not mean that groups with identical characteristics are selected in each specific situation of random assignment. Instead, it says something close to that: The probability of selecting a case can be mathematically determined, and, in the long run, the groups will be identical.

Sampling and random assignment are processes of systematically selecting cases for inclusion in a study. When a researcher randomly assigns, he or she sorts a collection of cases into two or more groups using a random process. By contrast, in random sampling, he or she selects a smaller subset of cases from a larger pool of cases (see Figure 8.1). A researcher can both sample and randomly assign. He or she can first sample to obtain a smaller set of cases (e.g., 150 people out of 20,000) and then use random assignment to divide the smaller set into groups (e.g., divide the 150 people into three groups of 50).

FIGURE 8.1 Random Assignment and Random Sampling



How to Randomly Assign

Random assignment is very simple in practice. A researcher begins with a collection of cases (individuals, organizations, or whatever the unit of analysis is), then divides it into two or more groups by a random process, such as asking people to count off, tossing a coin, or throwing dice. For example, a researcher wants to divide 32 people into two groups of 16. A random method is writing each person's name on a slip of paper, putting the slips in a hat, mixing the slips with eyes closed, then drawing the first 16 names for group 1 and the second 16 for group 2.

Matching versus Random Assignment

If the purpose of random assignment is to get two (or more) equivalent groups, would it not be simpler to match the characteristics of cases in each group? Some researchers match cases in groups on certain characteristics, such as age and sex. Matching is an alternative to random assignment, but it is an infrequently used one.

Matching presents a problem: What are the relevant characteristics to match on, and can one locate exact matches? Individual cases differ in thousands of ways, and the researcher cannot know which might be relevant. For example, a

researcher compares two groups of 15 students. There are 8 males in one group, which means there should be 8 males in the other group. Two males in the first group are only children; one is from a divorced family, one from an intact family. One is tall, slender, and Jewish; the other is short, heavy, and Methodist. In order to match groups, does the researcher have to find a tall Jewish male only child from a divorced home and a short Methodist male only child from an intact home? The tall, slender, Jewish male only child is 22 years old and is studying to become a physician. The short, heavy Methodist male is 20 years old and wants to be an accountant. Does the researcher also need to match the age and career aspirations of the two males? True matching soon becomes an impossible task.

EXPERIMENTAL DESIGN LOGIC

The Language of Experiments

Experimental research has its own language or set of terms and concepts. You already encountered the basic ideas: random assignment and independent and dependent variables. In experimental research, the cases or people used in research projects and on whom variables are measured are called the *subjects.*

Parts of the Experiment. We can divide the experiment into seven parts. Not all experiments have all these parts, and some have all seven parts plus others. The following seven, to be discussed here, make up a true experiment:

1. Treatment or independent variable
2. Dependent variable
3. Pretest
4. Posttest
5. Experimental group
6. Control group
7. Random assignment

In most experiments, a researcher creates a situation or enters into an ongoing situation, then modifies it. The *treatment* (or the stimulus or manipulation) is what the researcher modifies. The term comes from medicine, in which a physician administers a treatment to patients; the physician intervenes in a physical or psychological condition to change it. It is the independent variable or a combination of independent variables. In earlier examples of measurement, a researcher developed a measurement instrument or indicator (e.g., a survey question), then applied it to a person or case. In experiments, researchers "measure" independent variables by creating a condition or situation. For example, the independent variable is "degree of fear or anxiety"; the levels are high fear and low fear. Instead of asking subjects whether they are fearful, experimenters put subjects into either a high-fear or a low-fear situation. They measure the independent variable by manipulating conditions so that some subjects feel a lot of fear and others feel little.

Researchers go to great lengths to create treatments. Some are as minor as giving different groups of subjects different instructions. Others can be as complex as putting subjects into situations with elaborate equipment, staged physical settings, or contrived social situations to manipulate what the subjects see or feel. Researchers want the treatment to have an impact and produce specific reactions, feelings, or behaviors.

For example, a mock jury decision is one type of a treatment. Johnson (1985) asked subjects to watch a videotape of a child-abuse trial about a man who brought his 2-year-old son to an emergency room with a skull fracture. The videotapes were the same, except that in one, the man's attorney argued that the father was a highly religious person who followed the word of God in the Bible in all family affairs. In the other videotape, no such statement was made. The dependent variable was a decision of guilty or innocent and a recommended sentence for guilty decisions. Contrary to common sense, Johnson found that subjects were more likely to

find the religious defendant guilty and to recommend longer sentences.

Dependent variables or outcomes in experimental research are the physical conditions, social behaviors, attitudes, feelings, or beliefs of subjects that change in response to a treatment. Dependent variables can be measured by paper-and-pencil indicators, observation, interviews, or physiological responses (e.g., heartbeat or sweating palms).

Frequently, a researcher measures the dependent variable more than once during an experiment. The *pretest* is the measurement of the dependent variable prior to introduction of the treatment. The *posttest* is the measurement of the dependent variable after the treatment has been introduced into the experimental situation.

Experimental researchers often divide subjects into two or more groups for purposes of comparison. A simple experiment has two groups, only one of which receives the treatment. The *experimental group* is the group that receives the treatment or in which the treatment is present. The group that does not receive the treatment is called the *control group.* When the independent variable takes on many different values, more than one experimental group is used.

Steps in Conducting an Experiment. Following the basic steps of the research process, experimenters decide on a topic, narrow it into a testable research problem or question, then develop a hypothesis with variables. Once a researcher has the hypothesis, the steps of experimental research are clear.

A crucial early step is to plan a specific experimental design (to be discussed). The researcher decides the number of groups to use, how and when to create treatment conditions, the number of times to measure the dependent variable, and what the groups of subjects will experience from beginning to end. He or she also develops measures of the dependent variable and pilot tests the experiment (see Box 8.1).

BOX 8.1 Steps in Conducting an Experiment

1. Begin with a straightforward hypothesis that is appropriate for experimental research.
2. Decide on an experimental design that will test the hypothesis within practical limitations.
3. Decide how to introduce the treatment or create a situation that induces the independent variable.
4. Develop a valid and reliable measure of the dependent variable.
5. Set up an experimental setting and conduct a pilot test of the treatment and dependent variable measures.
6. Locate appropriate subjects or cases.
7. Randomly assign subjects to groups (if random assignment is used in the chosen research design) and give careful instructions.
8. Gather data for the pretest measure of the dependent variable for all groups (if a pretest is used in the chosen design).
9. Introduce the treatment to the experimental group only (or to relevant groups if there are multiple experimental groups) and monitor all groups.
10. Gather data for posttest measure of the dependent variable.
11. *Debrief* the subjects by informing them of the true purpose and reasons for the experiment. Ask subjects what they thought was occurring. Debriefing is crucial when subjects have been deceived about some aspect of the experiment.
12. Examine data collected and make comparisons between different groups. Where appropriate, use statistics and graphs to determine whether or not the hypothesis is supported.

The experiment itself begins after a researcher locates subjects and randomly assigns them to groups. Subjects are given precise, preplanned

instructions. Next, the researcher measures the dependent variable in a pretest before the treatment. One group is then exposed to the treatment. Finally, the researcher measures the dependent variable in a posttest. He or she also interviews subjects about the experiment before they leave. The researcher records measures of the dependent variable and examines the results for each group to see whether the hypothesis receives support.

Control in Experiments. Control is crucial in experimental research. A researcher wants to control all aspects of the experimental situation to isolate the effects of the treatment and eliminate alternative explanations. Aspects of an experimental situation that are not controlled by the researcher are alternatives to the treatment for change in the dependent variable and undermine his or her attempt to establish causality.

Experimental researchers use deception to control the experimental setting. *Deception* occurs when the researcher intentionally misleads subjects through written or verbal instructions, the actions of others, or aspects of the setting. It may involve the use of *confederates* or stooges—people who pretend to be other subjects or bystanders but who actually work for the researcher and deliberately mislead subjects. Through deception, the researcher tries to control what the subjects see and hear and what they believe is occurring. For example, a researcher's instructions falsely lead subjects to believe that they are participating in a study about group cooperation. In fact, the experiment is about male/female verbal interaction, and what subjects say is being secretly tape recorded. Deception lets the researcher control the subjects' definition of the situation. It prevents them from altering their cross-sex verbal behavior because they are unaware of the true research topic. By focusing their attention on a false topic, the researcher induces the unaware subjects to act "naturally." For realistic deception, researchers may invent false treatments and dependent variable measures to keep subjects unaware of the true ones. The use of deception in experiments raises ethical issues (to be discussed).

Types of Design

Researchers combine parts of an experiment (e.g., pretests, control groups, etc.) together into an *experimental design.* For example, some designs lack pretests, some do not have control groups, and others have many experimental groups. Certain widely used standard designs have names.

You should learn the standard designs for two reasons. First, in research reports, researchers give the name of a standard design instead of describing it. When reading reports, you will be able to understand the design of the experiment if you know the standard designs. Second, the standard designs illustrate common ways to combine design parts. You can use them for experiments you conduct or create your own variations.

The designs are illustrated with a simple example. A researcher wants to learn whether wait staff (waiters and waitresses) receive more in tips if they first introduce themselves by first name and return to ask "Is everything fine?" 8 to 10 minutes after delivering the food. The independent variable is the size of the tip received. The study occurs in two identical restaurants on different sides of a town that have had the same types of customers and average the same amount in tips.

Classical Experimental Design. All designs are variations of the *classical experimental design,* the type of design discussed so far, which has random assignment, a pretest and a posttest, an experimental group, and a control group.

Example. The experimenter gives 40 newly hired wait staff an identical two-hour training session and instructs them to follow a script in which they are not to introduce themselves by first name and not to return during the meal to check on the customers. They are next randomly divided into two equal groups of 20 and sent to the two restaurants to begin employment. The experimenter records the amount in tips for all

subjects for one month (pretest score). Next, the experimenter "retrains" the 20 subjects at restaurant 1 (experimental group). The experimenter instructs them henceforth to introduce themselves to customers by first name and to check on the customers, asking, "Is everything fine?" 8 to 10 minutes after delivering the food (treatment). The group at restaurant 2 (control group) is "retained" to continue without an introduction or checking during the meal. Over the second month, the amount of tips for both groups is recorded (posttest score).

Preexperimental Designs. Some designs lack random assignment and are compromises or shortcuts. These *preexperimental designs* are used in situations where it is difficult to use the classical design. They have weaknesses that make inferring a causal relationship more difficult.

One-Shot Case Study Design. Also called the one-group posttest-only design, the *one-shot case study design* has only one group, a treatment, and a posttest. Because there is only one group, there is no random assignment.

Example. The experimenter takes a group of 40 newly hired wait staff and gives all a two-hour training session in which they are instructed to introduce themselves to customers by first name and to check on the customers, asking, "Is everything fine?" 8 to 10 minutes after delivering the food (treatment). All subjects begin employment, and the experimenter records the amount in tips for all subjects for one month (posttest score).

One-Group Pretest-Posttest Design. This design has one group, a pretest, a treatment, and a posttest. It lacks a control group and random assignment.

Example. The experimenter takes a group of 40 newly hired wait staff and gives all a two-hour training session. They are instructed to follow a script in which they are not to introduce themselves by first name and not to return during the meal to check on the customers. All begin employment, and the experimenter records the amount in tips for all subjects for one month (pretest score). Next, the experimenter "retrains" all 40 subjects (experimental group). The experimenter instructs the subjects henceforth to introduce themselves to customers by first name and to check on the customers, asking, "Is everything fine?" 8 to 10 minutes after delivering the food (treatment). Over the second month, the amount of tips is recorded (posttest score).

This is an improvement over the one-shot case study because the researcher measures the dependent variable both before and after the treatment. But it lacks a control group. The researcher cannot know whether something other than the treatment occurred between the pretest and the posttest to cause the outcome.

Static Group Comparison. Also called the posttest-only nonequivalent group design, *static group comparison* has two groups, a posttest, and treatment. It lacks random assignment and a pretest. A weakness is that any posttest outcome difference between the groups could be due to group differences prior to the experiment instead of to the treatment.

Example. The experimenter gives 40 newly hired wait staff an identical two-hour training session and instructs them to follow a script in which they are not to introduce themselves by first name and not to return during the meal to check on the customers. They can choose one of the two restaurants to work at, as long as each restaurant ends up with 20 people. All begin employment. After one month, the experimenter "retrains" the 20 subjects at restaurant 1 (experimental group). The experimenter instructs them henceforth to introduce themselves to customers by first name and to check on the customers, asking, "Is everything fine?" 8 to 10 minutes after delivering the food (treatment). The group at restaurant 2 (control group) is

"retained" to continue without an introduction or checking during the meal. Over the second month, the amount of tips for both groups is recorded (posttest score).

Quasi-Experimental and Special Designs. These designs, like the classical design, make identifying a causal relationship more certain than do preexperimental designs. *Quasi-experimental designs* help researchers test for causal relationships in a variety of situations where the classical design is difficult or inappropriate. They are called *quasi* because they are variations of the classical experimental design. Some have randomization but lack a pretest, some use more than two groups, and others substitute many observations of one group over time for a control group. In general, the researcher has less control over the independent variable than in the classical design (see Table 8.1).

Two-Group Posttest-Only Design. This is identical to the static group comparison, with one exception: The groups are randomly assigned. It has all the parts of the classical design except a pretest. The random assignment reduces the chance that the groups differed before the treatment, but without a pretest, a researcher cannot be as certain that the groups began the same on the dependent variable. For example, Johnson and Johnson (1985) used a two-group posttest-only design. In the experiment, sixth-grade students were randomly assigned to one of two conditions: work groups in which points were awarded for how well the entire class learned material, or groups in which each group competed against others for points. All groups were mixed by race, sex, and ability level. Several dependent variables were measured, including academic achievement, cooperation across racial groups, and attitude toward others. The dependent variables were only measured after working in groups on an instruction unit for 10 days. The main result was that cooperative groups were more likely to promote cooperation and friendship across racial lines.

In another example study of a two-group posttest-only design with random assignment, Rind and Strohmetz (1999) conducted a study on messages about a upcoming special written on the back of customers' checks. The subjects were 81 dining parties eating at an upscale restaurant in New Jersey. The treatment was whether a female server wrote a message about an upcoming restaurant special on the back of a check and the dependent variable was the size of tips. The server with two years' experience was given a randomly shuffled stack of cards, half of which said No Message and half of which said Message. Just before she gave a customer his or her check, she randomly pulled a card from her pocket. If it said Message, she wrote about an up-

TABLE 8.1 A Comparison of the Classical Experimental Design with Other Major Designs

Design	Random Assignment	Pretest	Posttest	Control Group	Experimental Group
Classical	Yes	Yes	Yes	Yes	Yes
One-Shot Case Study	No	No	Yes	No	Yes
One-Group Pretest Postest	No	Yes	Yes	No	Yes
Static Group Comparison	No	No	Yes	Yes	Yes
Two-Group Posttest Only	Yes	No	Yes	Yes	Yes
Time Series Designs	No	Yes	Yes	No	Yes

coming special on the back of the customer's check. If it said No Message, she wrote nothing. The experimenters recorded the amount of the tip and the number of people at the table. They instructed the server to act the same toward all customers. The results showed that higher tips came from customers who received the message about upcoming specials.

Interrupted Time Series. In an *interrupted time series* design, a researcher uses one group and makes multiple pretest measures before and after the treatment. For example, after remaining level for many years, in 1990, cigarette taxes jumped 35 percent. Taxes remained relatively constant for the next 10 years. The hypothesis is that increases in taxes lower cigarette consumption. A researcher plots the rate of cigarette consumption for 1980 through 2000. The researcher notes that cigarette consumption was level during the 10 years prior to the new taxes, then dropped in 1990 and stayed about the same for the next 10 years.

Equivalent Time Series. An *equivalent time series* is another one-group design that extends over a time period. Instead of one treatment, it has a pretest, then a treatment and posttest, then treatment and posttest, then treatment and posttest, and so on. For example, people who drive motorcycles were not required to wear helmets before 1975, when a law was passed requiring helmets. In 1981, the law was repealed because of pressure from motorcycle clubs. The helmet law was reinstated in 1998. The researcher's hypothesis is that wearing protective helmets results in a lower number of head injury deaths in accidents. The researcher plots head injury death rates in motorcycle accidents over time. He or she finds the rate was very high prior to 1975, dropped sharply between 1975 and 1981, then rose to pre-1975 levels between 1981 and 1998, then dropped again from 1998 to the present.

Latin Square Designs. Researchers interested in how several treatments given in different sequences or time orders affect a dependent variable can use a *Latin square design.* For example, a geography instructor has three units to teach students: map reading, using a compass, and the longitude/latitude (LL) system. The units can be taught in any order, but the teacher wants to know which order most helps students learn. In one class, students first learn to read maps, then how to use a compass, then the LL system. In another class, using a compass comes first, then map reading, then the LL system. In a third class, the instructor first teaches the LL system, then compass usage, and ends with map reading. The teacher gives tests after each unit, and students take a comprehensive exam at the end of the term. The students were randomly assigned to classes, so the instructor can see whether presenting units in one sequence or another resulted in improved learning.

Solomon Four-Group Design. A researcher may believe that the pretest measure has an influence on the treatment or dependent variable. A pretest can sometimes sensitize subjects to the treatment or improve their performance on the posttest (see the discussion of testing effect to come). Richard L. Solomon developed the *Solomon four-group design* to address the issue of pretest effects. It combines the classical experimental design with the two-group posttest-only design and randomly assigns subjects to one of four groups. For example, a mental health worker wants to determine whether a new training method improves clients' coping skills. The worker measures coping skills with a 20-minute test of reactions to stressful events. Because the clients might learn coping skills from taking the test itself, a Solomon four-group design is used. The mental health worker randomly divides clients into four groups. Two groups receive the pretest; one of them gets the new training method and the other gets the old method. Another two groups receive no pretest; one of them gets the new method and the other the old method. All four groups are given the same posttest and the posttest results are compared.

If the two treatment (new method) groups have similar results, and the two control (old method) groups have similar results, then the mental health worker knows pretest learning is not a problem. If the two groups with a pretest (one treatment, one control) differ from the two groups without a pretest, then the worker concludes that the pretest itself may have an effect on the dependent variable.

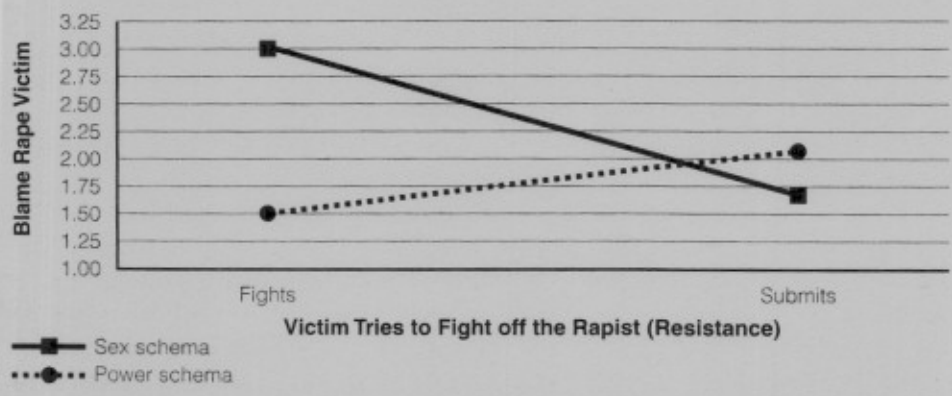
Factorial Designs. Sometimes, a research question suggests looking at the simultaneous effects of more than one independent variable. A *factorial design* uses two or more independent variables in combination. Every combination of the categories in variables (sometimes called *factors*) is examined. When each variable contains several categories, the number of combinations grows very quickly. The treatment or manipulation is not each independent variable; rather, it is each combination of the categories.

The treatments in a factorial design can have two kinds of effects on the dependent variable: main effects and interaction effects. Only *main effects* are present in one-factor or single-treatment designs. In a factorial design, specific combinations of independent variable categories can also have an effect. They are called *interaction effects* because the categories in a combination interact to produce an effect beyond that of each variable alone.

Interaction effects are illustrated in Figure 8.2, which uses data from a study by Ong and Ward (1999). As part of a study of 128 female undergraduates at the National University of Singapore, Ong and Ward measured which of two major ways subjects understood the crime of rape. Some of the women primarily understood it as sex and due to the male sex drive (sex schema); others understood it as primarily an act of male power and domination of a woman (power schema). The researchers asked the subjects to read a realistic scenario about the rape of a college student at their university. One randomly selected group of subjects read a scenario in which the victim tried to fight off the rapist. In the other set, she passively submitted. The researchers next asked the subjects to evaluate the degree to which the rape victim was at blame or responsible for the rape.

Results showed that the women who held the sex schema (and who also tended to embrace traditionalist gender role beliefs) more strongly blamed the victim when she resisted. Blame decreased if she submitted. The women who held a power schema (and who also tended to be nontraditionalists) were less likely to blame the victim if she fought. They blamed her more if she

FIGURE 8.2 Blame, Resistance, and Schema



passively submitted. Thus, the subjects' responses to the victim's act of resisting the attack varied by, or interacted with, their understanding of the crime of rape (i.e., the rape schema held by each subject). The researchers found that two rape schemas caused subjects to interpret victim resistance in opposite ways for the purpose of assigning responsibility for the crime.

Researchers discuss factorial design in a shorthand way. A "two by three factorial design" is written 2 × 3. It means that there are two treatments, with two categories in one and three categories in the other. A 2 × 3 × 3 design means that there are three independent variables, one with two categories and two with three categories each.

Valentine-French and Radtke (1989) used a 2 × 2 × 3 factorial design to study the effect of victim reaction to sexual harassment blame. The subjects were 120 male and 120 female undergraduate volunteers from the University of Calgary. The researchers operationalized the independent variable as an audiotaped vignette in which a professor guaranteed a good grade to a student if she or he was willing to cooperate, permitted caressing of the student's shoulder, and let the professor kiss her or him on the cheek. The experimenters varied the situation by having the student victim be male or female and by using one of three endings: The victim blamed his or her own behavior for the incident, blamed the professor, or gave no reaction. Thus, there were six combinations of victim gender and endings.

The subjects did not know the purpose of the study and listened to the vignette alone. The experimenters measured various background characteristics of the subjects with a questionnaire, as well as the main dependent variable—attribution of blame, or who was at fault. They operationalized the variable as an eight-item index measured with a 7-point Likert scale. Valentine-French and Radtke (1989) found that women were more likely to label the incident as sexual harassment and blame the professor. Male subjects, more than females, blamed the victim when the victim made a statement of self-blame. This was a 2 × 2 × 3 factorial design because three independent variables were examined: the subject's gender, the victim's gender, and the victim's reactions.

Design Notation

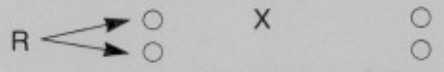
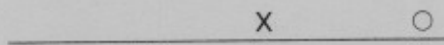
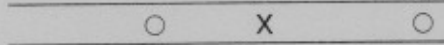
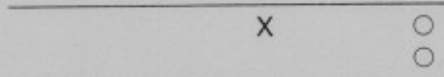
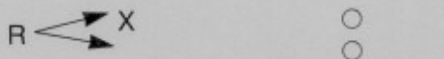
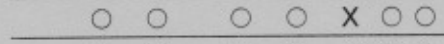
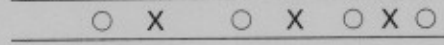
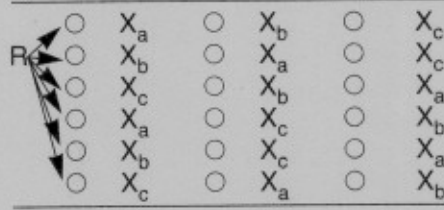
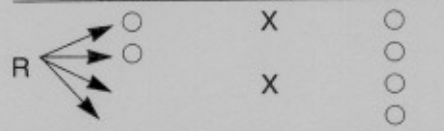
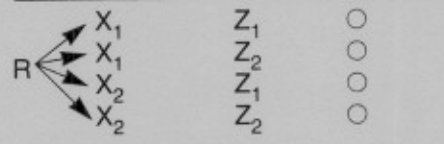
Experiments can be designed in many ways. *Design notation* is a shorthand system for symbolizing the parts of experimental design. Once you learn design notation, you will find it easier to think about and compare designs. For example, design notation expresses a complex, paragraph-long description of the parts of an experiment in five or six symbols arranged in two lines. It uses the following symbols: O = observation of dependent variable; X = treatment, independent variable; R = random assignment. The Os are numbered with subscripts from left to right based on time order. Pretests are O_1, posttests O_2. When the independent variable has more than two levels, the Xs are numbered with subscripts to distinguish among them. Symbols are in time order from left to right. The R is first, followed by the pretest, the treatment, and then the posttest. Symbols are arranged in rows, with each row representing a group of subjects. For example, an experiment with three groups has an R (if random assignment is used), followed by three rows of Os and Xs. The rows are on top of each other because the pretests, treatment, and posttest occur in each group at about the same time. Table 8.2 gives the notation for many standard experimental designs.

INTERNAL AND EXTERNAL VALIDITY

The Logic of Internal Validity

Internal validity means the ability to eliminate alternative explanations of the dependent variable.Variables, other than the treatment, that

TABLE 8.2 Summary of Experimental Designs with Notation

Name of Design	Design Notation
Classical experimental design	
Preexperimental Designs	
One-shot case study	
One-group pretest-posttest	
Static group comparison	
Quasi-Experimental Designs	
Two-group posttest only	
Interrupted time series	
Equivalent time series	
Latin square designs	
Solomon four-group design	
Factorial designs	

affect the dependent variable are threats to internal validity. They threaten the researcher's ability to say that the treatment was the true causal factor producing change in the dependent variable. Thus, the logic of internal validity is to rule out variables other than the treatment by controlling experimental conditions and through experimental designs. Next, we examine major threats to internal validity.

Threats to Internal Validity

The following are nine common threats to internal validity.[1]

Selection Bias. *Selection bias* is the threat that research participants will not form equivalent groups. It is a problem in designs without random assignment. It occurs when subjects in one

experimental group have a characteristic that affects the dependent variable. For example, in an experiment on physical aggressiveness, the treatment group unintentionally contains subjects who are football, rugby, and hockey players, whereas the control group is made up of musicians, chess players, and painters. Another example is an experiment on the ability of people to dodge heavy traffic. All subjects assigned to one group come from rural areas, and all subjects in the other grew up in large cities. An examination of pretest scores helps a researcher detect this threat, because no group differences are expected.

History. This is the threat that an event unrelated to the treatment will occur during the experiment and influence the dependent variable. *History effects* are more likely in experiments that continue over a long time period. For example, halfway through a two-week experiment to evaluate subjects' attitudes toward space travel, a spacecraft explodes on the launch pad, killing the astronauts. The history effect can occur in the cigarette tax example discussed earlier (see the discussion of interrupted time-series design). If a public antismoking campaign or reduced cigarette advertising also began in 1989, it would be hard to say that higher taxes caused less smoking.

Maturation. This is the threat that some biological, psychological, or emotional process within the subjects and separate from the treatment will change over time. *Maturation* is more common in experiments over long time periods. For example, during an experiment on reasoning ability, subjects become bored and sleepy and, as a result, score lower. Another example is an experiment on the styles of children's play between grades 1 and 6. Play styles are affected by physical, emotional, and maturation changes that occur as the children grow older, instead of or in addition to the effects of a treatment. Designs with a pretest and control group help researchers determine whether maturation or history effects are present, because both experimental and control groups will show similar changes over time.

Testing. Sometimes, the pretest measure itself affects an experiment. This *testing effect* threatens internal validity because more than the treatment alone affects the dependent variable. The Solomon four-group design helps a researcher detect testing effects. For example, a researcher gives students an examination on the first day of class. The course is the treatment. He or she tests learning by giving the same exam on the last day of class. If subjects remember the pretest questions and this affects what they learned (i.e., paid attention to) or how they answered questions on the posttest, a testing effect is present. If testing effects occur, a researcher cannot say that the treatment alone has affected the dependent variable.

Instrumentation. This threat is related to reliability. It occurs when the *instrument* or dependent variable measure changes during the experiment. For example, in a weight-loss experiment, the springs on the scale weaken during the experiment, giving lower readings in the posttest. Another example might have occurred in an experiment by Bond and Anderson (1987) on the reluctance to transmit bad news. The experimenters asked subjects to tell another person the results of an intelligence test and varied the test results to be either well above or well below average. The dependent variable was the length of time it took to tell the test taker the results. Some subjects were told that the session was being videotaped. During the experiment, the video equipment failed to work for one subject. If it had failed to work for more than one subject or had worked for only part of the session, the experiment would have had instrumentation problems. (By the way, subjects took longer to deliver bad news only if they thought they were doing so publicly—that is, being videotaped.)

Mortality. *Mortality,* or attrition, arises when some subjects do not continue throughout the

experiment. Although the word *mortality* means death, it does not necessarily mean that subjects have died. If a subset of subjects leaves partway through an experiment, a researcher cannot know whether the results would have been different had the subjects stayed. For example, a researcher begins a weight-loss program with 50 subjects. At the end of the program, 30 remain, each of whom lost 5 pounds with no side effects. The 20 who left could have differed from the 30 who stayed, changing the results. Maybe the program was effective for those who left, and they withdrew after losing 25 pounds. Or perhaps the program made subjects sick and forced them to quit. Researchers should notice and report the number of subjects in each group during pretests and posttests to detect this threat to internal validity.

Statistical Regression. *Statistical regression* is not easy to grasp intuitively. It is a problem of extreme values or a tendency for random errors to move group results toward the average. It can occur in two ways.

One situation arises when subjects are unusual with regard to the dependent variable. Because they begin as unusual or extreme, subjects are unlikely to respond further in the same direction. For example, a researcher wants to see whether violent films make people act violently. He or she chooses a group of violent criminals from a high-security prison, gives them a pretest, shows violent films, then administers a posttest. To the researcher's shock, the prisoners are slightly less violent after the film, whereas a control group of prisoners who did not see the film are slightly more violent than before. Because the violent criminals began at an extreme, it is unlikely that a treatment could make them more violent; by random chance alone, they appear less extreme when measured a second time.[2]

A second situation involves a problem with the measurement instrument. If many research participants score very high (at the ceiling) or very low (at the floor) on a variable, random chance alone will produce a change between the pretest and the posttest. For example, a researcher gives 80 subjects a test, and 75 get perfect scores. He or she then gives a treatment to raise scores. Because so many subjects already had perfect scores, random errors will reduce the group average because those who got perfect scores can randomly move in only one direction—to get some answers wrong. An examination of scores on pretests will help researchers detect this threat to internal validity.

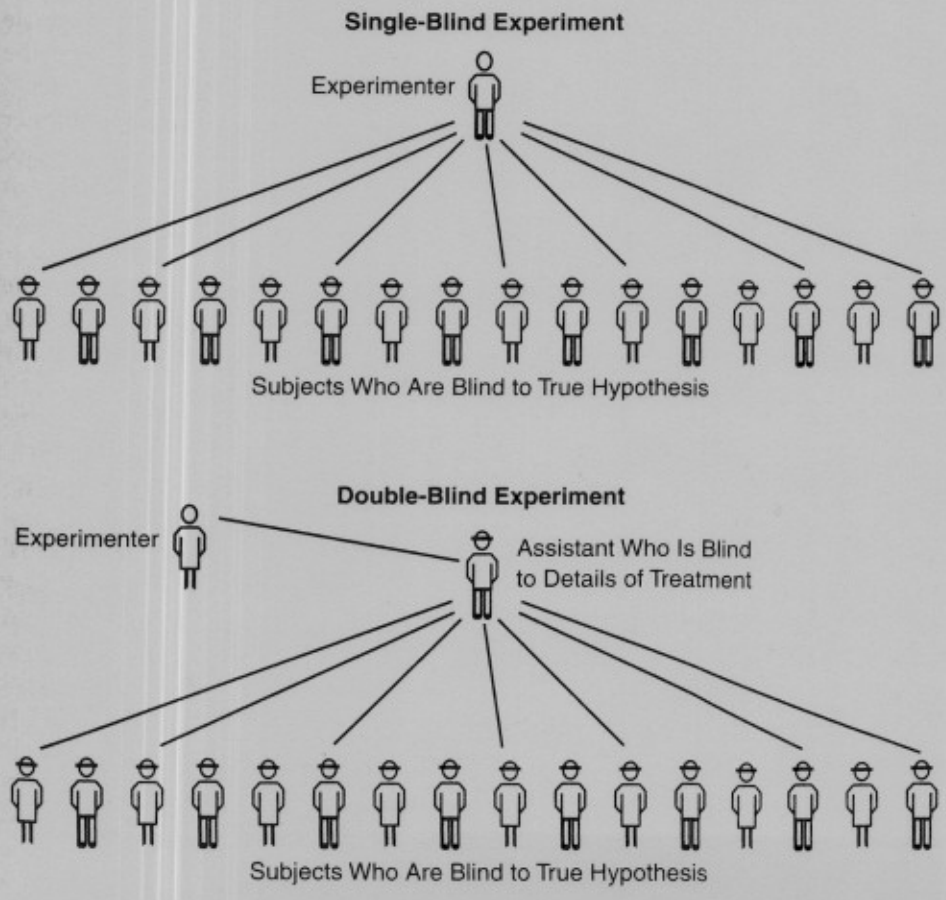
Diffusion of Treatment or Contamination. *Diffusion of treatment* is the threat that research participants in different groups will communicate with each other and learn about the other's treatment. Researchers avoid it by isolating groups or having subjects promise not to reveal anything to others who will become subjects. For example, subjects participate in a day-long experiment on a new way to memorize words. During a break, treatment group subjects tell those in the control group about the new way to memorize, which control group subjects then use. A researcher needs outside information such as postexperiment interviews with subjects to detect this threat.

Experimenter Expectancy. Although it is not always considered a traditional internal validity problem, the experimenter's behavior, too, can threaten causal logic.[3] A researcher may threaten internal validity, not by purposefully unethical behavior but by indirectly communicating *experimenter expectancy* to subjects. Researchers may be highly committed to the hypothesis and indirectly communicate desired findings to subjects. For example, a researcher studies the effects of memorization training on student learning ability, and also sees the grade transcripts of subjects. The researcher believes that students with higher grades tend to do better at the training and will learn more. Through eye contact, tone of voice, pauses, and other nonverbal communication, the researcher unconsciously trains the students with higher grades more intensely; the researcher's nonverbal behavior is the opposite for students with lower grades.

Here is a way to detect experimenter expectancy. A researcher hires assistants and teaches them experimental techniques. The assistants train subjects and test their learning ability. The researcher gives the assistants fake transcripts and records showing that subjects in one group are honor students and the others are failing, although in fact the subjects are identical. Experimenter expectancy is present if the fake honor students, as a group, do much better than the fake failing students.

The *double-blind experiment* is designed to control researcher expectancy. In it, people who have direct contact with subjects do not know the details of the hypothesis or the treatment. It is *double* blind because both the subjects and those in contact with them are blind to details of the experiment (see Figure 8.3). For example, a researcher wants to see if a new drug is effective. Using pills of three colors—green, yellow, and pink—the researcher puts the new drug in the yellow pill, puts an old drug in the pink one, and makes the green pill a *placebo*—a false treatment that appears to be real (e.g., a sugar pill without any physical effects). Assistants who give the pills and record the effects do not know which color

FIGURE 8.3 Double-Blind Experiments: An Illustration of Single-Blind, or Ordinary, and Double-Blind Experiments



contains the new drug. Only another person who does not deal with subjects directly knows which colored pill contains the drug and examines the results.

External Validity and Field Experiments

Even if an experimenter eliminates all concerns about internal validity, external validity remains a potential problem. *External validity* is the ability to generalize experimental findings to events and settings outside the experiment itself. If a study lacks external validity, its findings hold true only in experiments, making them useless to both basic and applied science.

Reactivity. Research participants might react differently in an experiment than they would in real life because they know they are in a study; this is called *reactivity.* The *Hawthorne effect* is a specific kind of reactivity.[4] The name comes from a series of experiments by Elton Mayo at the Hawthorne, Illinois, plant of Westinghouse Electric during the 1920s and 1930s. Researchers modified many aspects of working conditions (e.g., lighting, time for breaks, etc.) and measured productivity. They discovered that productivity rose after each modification, no matter what it was. This curious result occurred because the workers did not respond to the treatment but to the additional attention they received from being part of the experiment and knowing that they were being watched. Later research questioned whether this occurred, but the name is used for an effect from the attention of researchers. A related effect is the effect of something new, which may wear off over time.

Field Experiments. So far, this chapter has focused on experiments conducted under the controlled conditions of a laboratory. Experiments are also conducted in real-life or field settings where a researcher has less control over the experimental conditions. The amount of control varies on a continuum. At one end is the highly controlled *laboratory experiment,* which takes place in a specialized setting or laboratory; at the opposite end is the *field experiment,* which takes place in the "field"—in natural settings such as a subway car, a liquor store, or a public sidewalk. Subjects in field experiments are usually unaware that they are involved in an experiment and react in a natural way. For example, researchers have had a confederate fake a heart attack on a subway car to see how the bystanders react.[5]

A dramatic example is a field experiment by Harari and colleagues (1985) on whether a male passerby will attempt to stop an attempted rape. In this experiment, conducted at San Diego State University, an attempted rape was staged on a somewhat isolated campus path in the evening. The staged attack was clearly visible to unsuspecting male subjects who approached alone or in groups of two or three. In the attack, a female student was grabbed by a large man hiding in the bushes. As the man pulled her away and tried to cover her mouth, the woman dropped her books. She struggled and screamed. "No, no! Help, help, please help me!" and "Rape!" Hidden observers told the actors when to begin to stage the attack and noted the actions of subjects. Assistance was measured as movement toward the attack site or movement toward a police officer visible across a nearby parking lot. The study found that 85 percent of men in groups and 65 percent of men walking alone made a detectable move to assist the woman.

The amount of experimenter control is related to internal and external validity. Laboratory experiments tend to have greater internal validity but lower external validity; that is, they are logically tighter and better controlled, but less generalizable. Field experiments tend to have greater external validity but lower internal validity; that is, they are more generalizable but less controlled. Quasi-experimental designs are common in field experiments. For example, in the experiment involving the staged attempted rape, the experimenters recreated a very realistic situation with high external validity. It had more

external validity than putting people in a laboratory setting and asking them what they would do hypothetically. Yet, subjects were not randomly assigned. Any man who happened to walk by became a subject. The experimenters could not precisely control what the subject heard or saw. The measurement of subject response was based on hidden observers who may have missed some subject responses. Table 8.3 summarizes threats to internal and external validity.

PRACTICAL CONSIDERATIONS

Every research technique has informal tricks of the trade. They are pragmatic and based on common sense but account for the difference between the successful research projects of an experienced researcher and the difficulties a novice researcher faces. Three are discussed here.

Planning and Pilot Tests

All social research requires planning, and most quantitative researchers use pilot tests. During the planning phase of experimental research, a researcher thinks of alternative explanations or threats to internal validity and how to avoid them. The researcher also develops a neat and well-organized system for recording data. In addition, he or she should devote serious effort to pilot testing any apparatus (e.g., computers, video cameras, tape recorders, etc.) that will be used in the treatment situation, and he or she must train and pilot test confederates. After the pilot tests, the researcher should interview the pilot subjects to uncover aspects of the experiment that need refinement.

TABLE 8.3 Major Internal and External Validity Concerns

Internal Validity	External Validity and Reactivity
Selection bias	Hawthorne effect
History effect	
Maturation	
Testing	
Instrumentation	
Experimental mortality	
Statistical regression	
Diffusion of treatment	
Experimenter expectancy	

Instructions to Subjects

Most experiments involve giving instructions to subjects to set the stage. A researcher should word instructions carefully and follow a prepared script so that all subjects hear the same thing. This ensures reliability. The instructions are also important in creating a realistic cover story when deception is used.

Postexperiment Interview

At the end of an experiment, the researcher should interview subjects, for three reasons. First, if deception was used, the researcher needs to debrief the research participants, telling them the true purpose of the experiment and answering questions. Second, he or she can learn what the subjects thought and how their definitions of the situation affected their behavior. Finally, he or she can explain the importance of not revealing the true nature of the experiment to other potential participants.

RESULTS OF EXPERIMENTAL RESEARCH: MAKING COMPARISONS

Comparison is the key to all research. By carefully examining the results of experimental research, a researcher can learn a great deal about threats to internal validity, and whether the treatment has an impact on the dependent variable. For example, in

the Bond and Anderson (1987) experiment on delivering bad news, discussed earlier, it took an average of 89.6 and 73.1 seconds to deliver favorable versus 72.5 or 147.2 seconds to deliver unfavorable test scores in private or public settings, respectively. A comparison shows that delivering bad news in public takes the longest, whereas good news takes a bit longer in private.

A more complex illustration of such comparisons is shown in Figure 8.4 on the results of a series of five weight-loss experiments using the classical experimental design. In the example, the 30 research participants in the experimental group at Enrique's Slim Clinic lost an average of 50 pounds, whereas the 30 in the control group did not lose a single pound. Only one person dropped out during the experiment. Susan's Scientific Diet Plan had equally dramatic results, but 11 people in her experimental group dropped out. This suggests a problem with experimental mortality. People in the experimental group at Carl's Calorie Counters lost 8 pounds, compared to 2 pounds for the control group, but the control group and the experimental group began with an average of 31 pounds difference in weight. This suggests a problem with selection bias. Natalie's Nutrition Center had no experimental mortality or selection bias problems, but those in the experimental group lost no more weight than those in the control group. It appears that the treatment was not effective. Pauline's Pounds Off also avoided selection bias and experimental mortality problems. People in her experimental group lost 32 pounds, but so did those in the control group. This suggests that the maturation, history, or diffusion of treatment effects may have occurred. Thus, the treatment at Enrique's Slim Clinic appears to be the most effective one. See Box 8.2 for a practical application of comparing experimental results.

A WORD ON ETHICS

Ethical considerations are a significant issue in experimental research because experimental

FIGURE 8.4 Comparisons of Results, Classical Experimental Design, Weight-Loss Experiments

	Enrique's Slim Clinic	
	Pretest	*Posttest*
Experimental	190 (30)	140 (29)
Control group	189 (30)	189 (30)

	Natalie's Nutrition Center	
	Pretest	*Posttest*
Experimental	190 (30)	188 (29)
Control group	192 (29)	190 (28)

	Susan's Scientific Diet Plan	
	Pretest	*Posttest*
Experimental	190 (30)	141 (19)
Control group	189 (30)	189 (28)

	Pauline's Pounds Off	
	Pretest	*Posttest*
Experimental	190 (30)	158 (30)
Control group	191 (29)	159 (28)

	Carl's Calorie Counters	
	Pretest	*Posttest*
Experimental	160 (30)	152 (29)
Control group	191 (29)	189 (29)

BOX 8.2 A "Natural" Experiment on Law Compliance in New Orleans

Occasionally, a "natural" experiment in the field happens due to public policy changes or a government or other organizational intervention, and researchers are able to measure, participate, and learn from it. This greatly increases the *external validity* of an experiment. For example, until the mid-1990s, laws on selling liquor to underage customers were barely enforced in New Orleans, Louisiana. If caught, an offending liquor retailer met privately with the liquor commission and might pay a small fine. Enforcing liquor laws was low priority for state and local government, so only three enforcement officers were assigned to monitor 5,000 alcohol outlets in the New Orleans area. Public officials planned to shift enforcement priorities and Scribner and Cohen (2001) were able to examine its impact. The experimenters had several people who clearly looked under 18 years old attempt to purchase alcoholic beverages illegally (the law required being at least 21 years of age) at 143 randomly selected liquor outlets from November 1995 through January 1996 (Time 0). The percentage of them who could buy liquor illegally was the *pretest measure*. After assessing the rate of illegal sales, the *dependent variable*, the police issued citations to 51 of the sales outlets, the primary *independent variable* or treatment. Government officials initiated a media campaign urging better law compliance. There were *two posttest measures*, first in March to April 1996 (Time 1) and again in November 1996 to January 1997 (Time 2), during which the experimenters checked the 143 outlets.

DEPENDENT VARIABLE: PERCENTAGE OBEYING THE LAW (REFUSING TO SELL ILLEGALLY)

	Pretest (Time 0)	Posttest 1 (Time 1)	Posttest 2 (Time 2)	No. of Retail Liquor Outlets
Experimental	6.7%	51%	29%	45
Control	13.3%	35%	17%	98
Total	11.1%	40%	21%	143

The researchers' results allow us to compare rates of illegal selling activity before and after citations/media campaign (*pretest* and *posttest* measures) and to compare outlets that directly received (*experimental group*) citations with those that did not receive citations and only had greater media exposure (*control group*). By making comparisons among the results, we can see that the citations and campaign did not stop the illegal activity, but it had some effect. The impact was greater on outlets that had been directly punished. In addition, by adding a later follow-up, (Time 2), we see how the law-enforcement impact slowly decayed over time.

As often occurs in natural experiments, internal validity is threatened: First, the pretest measure shows a difference in the two sets of outlets, with outlets who received the treatment showing higher rates of illegal behavior; this is potential *selection bias*. Second, the media campaign occurred for all outlets, so the treatment is really a citation plus the media campaign. The authors note that they had intended to compare the New Orleans area with another area with neither the media or citation campaign, but were unable to do so. Since outlets that did not receive the treatment (i.e., a citation for law violation) probably learned about it from others in the same business, a form of *diffusion of the treatment* could be operating. Third, the researchers report that they began with 155 outlets, but only studied 143 because 12 outlets went out of business during the study. The authors note that none of the outlets that stopped selling alcohol closed due to new law enforcement, but if those outlets that received citations had more problems and were more

BOX 8.2 Continued

likely to go out of business, if could indicate *experimental mortality.* The experimenters do not mention any external events in New Orleans that happened during the time of the study (e.g., a publicized event such as underage drinker dying of alcohol poisoning from overdrinking). Researchers need to be aware of potential external events when a study continues for so long and consider possible *history effects.*

research is intrusive (i.e., it interferes). Treatments may involve placing people in contrived social settings and manipulating their feelings or behaviors. Dependent variables may be what subjects say or do. The amount and type of intrusion is limited by ethical standards. Researchers must be very careful if they place research participants in physical danger or in embarrassing or anxiety-inducing situations. They must painstakingly monitor events and control what occurs.

Deception is common in social experiments, but it involves misleading or lying to subjects. Such dishonesty is not condoned as acceptable and is acceptable only as the means to achieve a goal that cannot be achieved otherwise. Even for a worthy goal, deception can be used only with restrictions. The amount and type of deception should not go beyond what is minimally necessary, and research participants should be debriefed.

CONCLUSION

In this chapter, you learned about random assignment and the methods of experimental research. Random assignment is an effective way to create two (or more) groups, which can be treated as equivalent and hence compared. In general, experimental research provides precise and relatively unambiguous evidence for a causal relationship. It follows the positivist approach, produces quantitative results that can be analyzed with statistics, and is often used in evaluation research (see Box 8.2).

This chapter also examined the parts of an experiment and how they can be combined to produce different experimental designs. In addition to the classical experimental design, you learned about preexperimental and quasi-experimental designs. You also learned how to express them using design notation.

You learned that internal validity—the internal logical rigor of an experiment—is a key idea in experimental research. Threats to internal validity are possible alternative explanations to the treatment. You also learned about external validity and how field experiments maximize external validity.

The real strength of experimental research is its control and logical rigor in establishing evidence for causality. In general, experiments tend to be easier to replicate, less expensive, and less time consuming than the other techniques. Experimental research also has limitations. First, some questions cannot be addressed using experimental methods because control and experimental manipulation are impossible. Another limitation is that experiments usually test one or a few hypotheses at a time. This fragments knowledge and makes it necessary to synthesize results across many research reports. External validity is another potential problem because many experiments rely on small nonrandom samples of college students.[6]

You learned how a careful examination and comparison of results can alert you to potential

problems in research design. Finally, you saw some practical and ethical considerations in experiments.

In the next chapters, you will examine other research techniques. The logic of the nonexperimental methods differs from that of the experiment. Experimenters focus narrowly on a few hypotheses. They usually have one or two independent variables, a single dependent variable, a few small groups of subjects, and an independent variable that the researcher induces. By contrast, other social researchers test many hypotheses at once. For example, survey researchers measure a large number of independent and dependent variables and use a larger number of randomly sampled subjects. Their independent variables are usually preexisting conditions in research participants.

Key Terms

classical experimental design
control group
debrief
deception
demand characteristics
design notation
diffusion of treatment
double-blind experiment
equivalent time series
experimental design
experimental group
factorial design
field experiment
Hawthorne effect
history effects
interaction effect
interrupted time series
laboratory experiment
Latin square design
maturation
mortality
one-shot case study
placebo
posttest
preexperimental designs
pretest
quasi-experimental designs
random assignment
reactivity
selection bias
Solomon four-group design
static group comparison
subjects
treatment

Endnotes

1. For additional discussions of threats to internal validity, see Cook and Campbell (1979:51–68), Kercher (1992), Smith and Glass (1987), Spector (1981:24–27), and Suls and Rosnow (1988).
2. This example is borrowed from Mitchell and Jolley (1988:97).
3. Experimenter expectancy is discussed in Aronson and Carlsmith (1968:66–70), Dooley (1984:151–153), and Mitchell and Jolley (1988:327–329).
4. The Hawthorne effect is described in Roethlisberger and Dickenson (1939), Franke and Kaul (1978), and Lang (1992). Also see the discussion in Cook and Campbell (1979:123–125) and Dooley (1984:155–156). Gillespie (1988, 1991) discussed the political context of the experiments.
5. See Piliavin and associates (1969).
6. See Graham (1992) and Sears (1986).