Experimental Research

The experiment is distinguished by the activity of the researcher who determines the conditions under which investigation will take place. Wholly or in part, the researcher . . . creates, builds or controls the research setting.
— Willer and Walker, *Building Experiments, Testing Social Theory,* p. 2

Pager (2007) wanted to examine the impact of imprisonment on the chances of getting a job after release. In addition, he was curious about whether race had an effect. He created a field experiment in which he hired college-age male "testers." Half the testers were White and half were Black. In 2001, the testers applied for entry-level jobs that had been advertised in the newspaper in the Milwaukee metropolitan area. The jobs required no experience and only a high school diploma. Pager matched testers of each race on age, physical appearance, and presentation style. He trained the testers, checked their interview skills, and created a fake résumé for each. For one-half of the testers of each race, he created résumés that showed a felony conviction for drug possession and 18 months of prison time. The other half had a virtually identical résumé but no criminal record. Pager randomly assigned testers to the advertised jobs. In this study, the independent variables were tester race and criminal record. The dependent variable was whether an employer called back to offer a job to a tester. Pager found that testers with a criminal record on their résumé and the Black testers received far fewer job offers. When he looked at the two independent variables together, he learned that a White tester with a criminal record was more likely to be offered a job than an equally qualified Black tester who had no criminal record. In Wisconsin as in many other states, laws bar hiring discrimination by race and by criminal conviction when the conviction has no relevance to a job. Pager also looked at data suggesting the large racial effect he found in Milwaukee may be larger in other major urban areas.

This chapter will focus on research techniques that yield quantitative data. We begin with experiments.

Experimental research builds on the principles of a positivist approach.[1] Natural scientists (e.g., chemists or biologists) and researchers in related applied fields (e.g., agriculture, engineering, and medicine) conduct experiments. We use experiments in education, criminal justice, journalism, marketing, nursing, political science, psychology, social work, and sociology to examine many social issues and theories. As Pager's (2007) experiment on race and criminal record on job seeking in the opening box illustrates, the experiment provides us powerful evidence about how one or two variables affect a dependent variable.

In commonsense language, to *experiment* means to modify one thing in a situation and then compare 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 because I have a simple hypothesis that it is causing the problem. I try to start it again. I had modified one thing (cleaned the connections) and compared the outcome (whether the car started) to the previous situation (it did not start). An experiment begins with a "hypothesis about causes." My hypothesis was that a buildup of crud on the battery connections was blocking the flow of electricity and the cause of the car not starting, so once I had cleared off the crud, the car could start. This commonsense experiment is simple, but it illustrates three critical steps in an experiment: (1) start with causal hypothesis, (2) modify one specific aspect of a situation that is closely connected to the cause, and (3) compare outcomes.

In the chapter's opening box, Pager's (2007) hypothesis was that racial heritage and criminal record influence whether a qualified person will receive job offers. He selected testers by race and created false résumés to modify the job-seeking situation in ways connected to racial heritage and criminal record. He then compared the job offers by racial background and criminal record.

Compared to other social research techniques, experimental research offers the strongest tests of causal relationships. This is so because we consciously design an experiment to satisfy the three conditions for causality (i.e., temporal order in which the independent precedes the dependent variable, evidence of an association, and ruling out alternative causes).

APPROPRIATE TECHNIQUE

People new to social research may anguish over which research technique best fits a specific research question. It can be a difficult decision because there is no ready-made, fixed match between technique and question. Deciding requires making an "informed judgment." You can develop judgment skills by learning the strengths and weaknesses of the various research techniques, reading the methodology section of many published studies, assisting an experienced social researcher, and acquiring practical experience by conducting studies yourself.

An experiment can powerfully test and focus evidence about causal relationships. Compared to other research techniques, it has both advantages and limitations, and these help to see where it is most appropriate.

The experiment is often artificial. It is a purposeful simplification of the complex social world. We tend to think that "artificial" means something negative, but Webster and Sell (2007:11) argue,

> *The greatest benefits of experiments reside in the fact that they are artificial. That is, experiments allow observation in a situation that has been designed and created by investigators rather than one that occurs in nature.*

Artificial means that the experimenter consciously controls the study situation and purposely incorporates theoretically relevant variables while removing variables without a causal importance for a hypothesis. *Artificial* also means a sharpened focus and narrowly targeted effects that we may not easily encounter in the natural world. We include the independent and dependent variables, but exclude

irrelevant or **confounding variables** (i.e., variables not a part of our hypothesis test). An analogy is the chemist who finds pure sodium in the natural world. In a controlled laboratory setting, the chemist mixes it precisely with another pure chemical to study its effects. The controlled, sterile laboratory is artificial, pure sodium is artificial, and what the chemist mixes it with is artificial, yet the outcome can produce new knowledge and compounds that have great utility in the real world.

Social science experiments have a very powerful logic; however, we face many practical and ethical limitations. In an experiment, we manipulate some aspects of the world and then examine the outcomes; however, we cannot manipulate many areas of human life for the sake of gaining scientific knowledge. With experiments, we are limited to questions that have specific conditions that we can manipulate and that clearly fall within ethical standards for research with humans. Thus, an experiment cannot directly answer questions such as these: Do people who complete a college education increase their annual income more than people who do not attend college? 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? We cannot allow some people to attend college and prevent others from attending to discover who earns more income later in life. We cannot induce couples to have either many children or a single child in order to examine how leadership skills develop in the children. We cannot compel people to join or quit organizations or never join them and then see whether they vote. Although we cannot manipulate many of the situations or variables we find of interest, we are able to be creative in simulating such interventions or conditions.

The experimental technique is usually best for issues that have a narrow scope or scale. We can often assemble and conduct numerous experiments with limited resources in a short period yet still test theoretically significant hypotheses. For example, we could replicate a study like that of Niven (see Example Box 1, News Reports on Death Penalty Opinions) in less than a month and at very low cost.

In general, an experiment is suited for micro-level (e.g., individual psychological or small-group phenomena) more than for macro-level theoretical concerns. This is why social psychologists and political psychologists conduct experiments. Experiments cannot easily address questions that require consideration of conditions operating across an entire society or over many years.

Experiments encourage us to isolate and target one or a few causal variables. Despite the strength to demonstrate the causal effect of one or two variables, experiments are not effective if we want to consider dozens of variables simultaneously. It is rarely appropriate for questions requiring us to examine the impact of many of variables together or to assess conditions across a range of complex settings or numerous social groups.

Experiments provide focused tests of hypotheses with each experiment considering one or two variables in a specific setting. Knowledge advances slowly by compiling, comparing, and synthesizing the findings from numerous separate experiments. This strategy for building knowledge differs from that of other research techniques in which one study might examine fifteen to twenty variables simultaneously in a diverse range of social settings.

Convention also influences the research questions that best align with the experimental method. Researchers have created vast research literature on many topics by using the experimental method. This has facilitated rapid, smooth communication about those topics. It has also facilitated replicating past experiments with minor adjustments and precisely isolating the effects of specific variables. Expertise in experiments can be a limitation because researchers who specialize in such topics tend to expect everyone to use the experimental method. These researchers evaluate new studies by the standards of a good experiment and may more slowly accept and assimilate new knowledge coming from a nonexperimental study.

Confounding variables In experimental research, factors that are not part of the intended hypothesis being tested, but that have effects on variables of interest and threaten internal validity.

EXAMPLE BOX 1

News Reports on Death Penalty Opinions

Niven (2002) noted the overwhelming support (75–80 percent) in opinion polls for the death penalty among Americans in recent decades. However, if people have a choice between supporting the death penalty for a murder or a sentence of life imprisonment without parole (LIWP), their support for the death penalty drops by nearly one-half. Niven found that more than 90 percent of media stories on death penalty opinions report overwhelming public support for it, but very few stories report that many people would prefer LIWP as an alternative punishment for the same crimes. Niven hypothesized that support for the death penalty might change if people had exposure to media stories that told them about high levels of public support for the LIWP alternative. To test his hypothesis, he went to waiting areas in the Miami International Airport for more than a two-week period and recruited 564 participants for his study. He randomly assigned people to read one of three newspaper articles, which were his independent variable. One newspaper article told about overwhelming support for the death penalty, another reported public support for LIWP, and the third was unrelated to the death penalty issue and about airport expansion plans. He told respondents a cover story: that the study was about newspaper article writing style. Participants completed a questionnaire about the clarity and organization of the article to disguise the purpose of the experiment. He also had a section on political beliefs under the premise that he wanted to know whether people with different political beliefs reacted the same way to the article. This section included his dependent variable, three questions about determining support or opposition for the death penalty for the crime of murder, preference for the death penalty or LIWP, and an estimate as to whether more or fewer states would adopt the death penalty in the future. His results showed no differences on the death penalty questions between participants who read about overwhelming death penalty support and the control group that read about airport expansion. More than 80 percent of both groups supported the death penalty, a little over one-half preferred it to LIWP, and most thought more states would adopt the death penalty in the future. People who read about LIWP showed much less support for the death penalty (62 percent), preferred LIWP over the death penalty (by a 57 to 43 percent margin), and predicted that fewer states would have the death penalty in the future. Thus, Nevin found support for his hypothesis that media stories that report on public support for the death penalty only perpetuate public opinion for it over the LIWP alternative.

We also can conduct mixed experimental and nonexperimental methods in a study to expand understanding. For example, we want to study attitudes toward people in wheelchairs. We could survey a thousand people about their views on people in wheelchairs. We could conduct a field research study and observe how people react to us while we are in a wheelchair in real-life settings. We can also design an experiment in which we interact with others—sometimes while in a wheelchair and at other times standing or walking without a wheelchair and then noting how people respond to each situation. To best test theories and develop a fuller understanding, we combine knowledge from all types of studies (see Example Box 2, Experimental and Survey Methods to Test and Apply Identity Theory).

A SHORT HISTORY OF THE EXPERIMENT

The social sciences, starting with psychology, borrowed the experimental method from the natural sciences. Psychology did not fully embrace the experiment until after 1900.[2] Wilhelm M. Wundt (1832–1920), a German psychologist and physiologist, introduced the experimental method into psychology. During the late 1800s, Germany was the center of graduate education, and social scientists came from around the world to study there. Wundt established a laboratory for experimentation in psychology that became a model for social research. By 1900, universities in the United States and elsewhere established psychology laboratories to conduct experimental research. However, William

EXAMPLE BOX 2

Experimental and Survey Methods to Test and Apply Identity Theory

Transue (2007) combined experimental logic with survey research methods in one study and tested an abstract social science theory by applying it to a real public policy issue. His work contributed to a growing literature showing how a subtle emphasis on racial differences among Americans tends to accentuate divisions along racial lines regarding public issues.

According to social identity theory, we automatically categorize other people into in-groups (groups to which we belong) and out-groups (groups to which we do not belong). These groups form the basis of social boundaries and feelings of social distance from or closeness to other people. We also have multiple identities. A subset of the broader theory, self-categorization, says we recategorize others as members of in-groups or out-groups based on which of our identities is more active. Social boundaries and feelings of social distance depend on the most salient in-group. We feel closer to members of an in-group and farther from people in salient out-groups. *Priming* is a process by which something happens to activate a particular identity. Once activated, this identity tends to have greater influence over subsequent behavior or thinking. Once reminded of an identity (i.e., it has been primed) it moves to the forefront of how we think about ourselves and therefore influences our behavior.

In most past studies on social identity theory, researchers used laboratory experiments with small convenience samples of students and tested the effect of a temporary, artificially created identity on a contrived issue. Transue (2007) sought more external validity. To obtain it, he used a large random sample of adults, an actual social identity, and a real public policy issue. His study used a telephone survey of a random sample of 405 White U.S. citizens in the Minneapolis metropolitan area in summer 1998 relying on random-digit dialing. Transue considered two actual identities, race and nation. He built on past studies that showed racially prejudiced Whites who had been primed or reminded of their race to be more likely to think in racist ways when they voted. The real policy issue he examined was support for paying taxes for public schools.

For the independent variable, social identity, Transue asked randomly assigned subsets of survey respondents one of two questions: "How close do you feel to your ethnic or racial group?" or "How close do you feel to other Americans?" This question primed or raised awareness of an identity. Later in the survey, he asked randomly assigned subsets of two questions about paying school taxes, "to improve education in public schools" or "to improve opportunities for minorities." This was the main dependent variable. Tansue hypothesized that Whites who were primed about their racial identity would reject paying taxes to help minorities more than Whites who were primed about their American national identity. He also thought that Whites primed about an American national identity would more strongly support taxes for public schools generally than those primed about their racial identity.

Transue found that Whites primed with a racial identity and asked about helping minorities had the least amount of support for paying school taxes. The most support came from Whites primed with an American national identity and asked about helping public schools generally. Tensue also looked at the Whites who had identified more strongly with their racial-ethnic group and compared them with Whites having a weak or no racial identification. Consistent with social identity theory, he found that Whites with the strongest racial identity showed the most resistance to paying taxes to improve minority opportunities. In this study, a primed racial self-identity increased the salience of a person's racial in-group and heightened social boundaries associated with racial categories. A strong identity with one's racial in-group increased social distance for people in racial out-groups and lowered a desire to provide them with assistance.

James (1842–1910), a prominent philosopher and psychologist, did not use or embrace the experimental method. The experiment displaced a more philosophical, introspective, integrative approach in psychology that was closer to the interpretive social science approach.

From 1900 to 1950, social researchers elaborated on the experimental method until it became

entrenched in some areas. The experiment's appeal was its objective, unbiased, scientific approach to studying mental and social life in an era when the scientific study of social life was just gaining broad public acceptance. Four trends sped the expansion of experimental social research: the rise of behaviorism, the spread of quantification, the changes in research participants, and the method's practical applications. Let us briefly consider each trend.

1. *Behaviorism* is an approach in psychology founded by the American James B. Watson (1878–1958) and expanded by B. F. Skinner (1904–1990). It emphasizes creating precise measures of observable behavior or outward manifestations of inner mental life and advocates the experiment to conduct rigorous empirical tests of hypotheses.

2. *Quantification,* or measuring social phenomena with numbers, expanded between 1900 and 1950. Researchers conceptualized social constructs as quantified measures and jettisoned other nonquantifiable constructs (e.g., spirit, consciousness, will) from empirical research. An example is measuring mental ability by using the IQ test. Originally developed by Alfred Binet (1857–1911), a Frenchman, researchers translated the test into English and revised it by 1916. It soon had widespread use and appeal as a way to represent something as subjective as a person's mental ability with a single score and became an objective, scientific way to rank people. Between the years of 1921 and 1936, more than 5,000 articles were published on intelligence tests.[3] Many scaling and index techniques were developed in this period, and social researchers began to use applied statistics.

3. Over time, the people used as participants changed. Early social research reports contained the names of the specific individuals who participated in a study, and most were professional researchers. Later reports treated participants anonymously and reported only the results of their actions. Over time, there was a shift to use college students or schoolchildren as research participants. The relationship between a researcher and the people studied became more distant. Such distancing reflected a trend for the experimenters to be more detached, remote, and objective from the people under study. Researchers saw reducing emotional engagement with research participants in their studies as becoming more neutral or value-free and truly "scientific" in a positivist sense.

4. As researchers became aware of an experiment's practical applications, businesses, governments, health care facilities, and schools increasingly used experimental methods for applied purposes. For example, the U.S. Army adopted intelligence tests during World War I to sort thousands of soldiers into different military positions. The leader of the "scientific management" movement in factories, Frederick W. Taylor (1856–1915), advocated using experiments in factories. He worked with management to modify factory conditions as a way to increase worker productivity. In the 1920s, educational researchers conducted many experiments on teaching methods and the effect of class size on learning.

By the 1950s and 1960s, researchers became more concerned with possible sources of alternative explanations, or confounding variables, that might slip into experimental design. Researchers designed experiments to reduce such potential errors and increasingly used statistical procedures in data analysis. A turning point in the increasingly rigorous design of social science experiments was a book by Campbell and Stanley (1963), who defined basic designs and issues in experimental methods.

By the 1970s, researchers increasingly evaluated the methodological rigor of studies. A related trend was the increased use of deception and a corresponding rise in concern about ethical issues. For example, the now common practice of debriefing did not come into use until the 1960s.[4] Over the last three decades, the trend has been to use more sophisticated experimental designs and statistical techniques for data analysis.

Experiments and Theory

We conduct two types of social science experiments: empirically based and theory-directed (see Willer and Walker, 2007a, 2007b). The practical process of doing an experiment differs little, but each type has different purposes. Most studies are empirically based.

In the empirically based experiment, our goal is to determine whether an independent variable has a significant effect on a specific dependent variable. We want to document and describe an effect (i.e., its size, direction, or form). Often we empirically demonstrate the effect in a controlled setting from which we can generalize to "real-life" conditions (see the discussion of external validity later in this chapter). We generalize our findings to natural or "real-world" settings. For example, Solomon Asch's (1955) famous experiment demonstrated the effect of conformity to group pressure by having eight students look at three lines. Once Asch demonstrated the power of group conformity, we generalized its effects beyond his specific study of eight students looking at three lines to many sizes of groups of all types of people engaged in most real-life tasks. The study by Pager (2007) that opened this chapter was an empirically based study. It demonstrated the effects of race and a criminal record on job seeking, as did the study by Niven (2002) on news reports and death penalty opinions (see Example Box 1). Niven's study demonstrated the effect of reading news reports on death penalty opinions.

In a theory-directed experiment, we proceed deductively by converting an abstract model of how we believe the world operates (i.e., theory) into a specific study design with specific measures. The experiment is a replica of the theoretical model. When we generalize from a theory-directed experiment, we generalize the theory as a model of how the world operates. Our primary task is to test the theory and learn whether there is empirical evidence for it. We are not concerned with finding a large effect of the independent variable; rather, we are concerned with finding that a theory's specific expectations or predictions closely match empirical findings. We worry less whether the experimental test of theory is highly artificial and nonrealistic to the natural world. Our primary concern is whether the empirical results match our theory. We seek many replication experiments to show repeatedly that the evidence matches the theory or that the theory can survive numerous tests. Indeed, as Webster and Sell (2007:21) argue, "experimental results themselves are really not interesting except as they bear on a theory."

We often use statistical techniques in experiments to see how likely the result predicted by the theory occurs. If the theory-predicted outcome has a low probability but occurs regularly, our confidence in the theory's correctness grows. Here is a simple example. My friend believes he can tell the difference between five brands of diet colas. I have him drink twenty cups of them over 4 days. One-fifth of the cups is one brand and their order is totally mixed. If he is correct twenty of the twenty times, I am confident that he really can tell the difference. By chance alone, he would be correct only 20 percent of the time. If a theory such as the one regarding my friend is correct 100 percent of the time, our confidence in it grows, but 100 percent is rare. However, if my friend was correct 90 percent of the time, I would think his evaluation was very good but not perfect. If he was correct just 30 percent of the time, this is little better than chance alone, so my confidence in his evaluation is low. In theory testing, our confidence in an explanation varies by whether the theory's predictions far exceed what we expect by chance alone and whether it survives repeated tests.

The study by Transue reported in Example Box 2 has features of a theory-directed experiment. He sought to replicate tests of a theory that had survived many previous experimental tests, self-categorization theory. He applied the priming effect to activate self-categorization to select an in-group identity and then provided evidence that supported the theory. His study was unusual in that it combined survey methods and a realistic policy issue. Another study on the contact hypothesis described later in this chapter (see Example Box 7, A Field Experiment on College Roommates) is also a theory-directed experiment, although applied in a real-life situation. Although we usually begin theory-directed experiments in highly controlled artificial settings, we may extend and replicate them in naturalistic settings.

RANDOM ASSIGNMENT

As researchers, we are always making comparisons. 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 what is fundamentally alike.

There are many ways to compare.[5] We can compare the same person over time (e.g., before and after completing a training course)—a within-subject experiment. However, we are often less interested whether a treatment or independent variable results in one person changing than whether it generally has an effect. We can compare a group of people at two times (e.g., the group average of thirty people before and after a training course). We can also compare the same group of thirty people over a series treatments (e.g., three training programs in sequence) to see whether each time we get an effect. These are within-group experiments. Alternatively, we can also compare two groups of fifteen participants: fifteen who have had and another fifteen who have not had the treatment (e.g., the training course). This is a between-group experiment.

Random assignment facilitates between-group comparisons by creating similar groups. For comparative purposes, we do not want the group to differ with regard to variables that may present alternative explanations for a causal relationship. For example, we want to compare two groups to determine the causal effect of completing a firefighting training course on each person's ability to respond to a fire. We want the two groups to be similar in all respects except for taking the course. If the groups were identical except for the course, we can compare outcomes with confidence and know that the course caused any of the differences we found. If the groups differed (e.g., one had experienced firefighters or one had much younger and more physically fit participants) we could not be certain when we compared them that the training course was the only cause of any differences we observe.

Why Assign Randomly

Random assignment is a method for assigning cases (e.g., individuals, organizations) to groups to make comparisons. It is a way to divide a collection of participants into two or more groups to increase your confidence that the groups do not differ in a systematic way. It is a purely mechanical method; the assignment is automatic. You cannot assign based on your or a participant's personal preference or his or her features (e.g., you thought the person acted friendly, someone wants to be in a group with a friend, put all people who arrived late in one group).

Random assignment Participants divided into groups at the beginning of experimental research using a random process so the experimenter can treat the groups as equivalent.

Random assignment is random in a statistical or mathematical sense, not in an everyday sense. We may say *random* to mean unplanned, haphazard, or accidental. In probability theory, *random* is a process in which each case has an equal chance of being selected. With random selection, you can mathematically calculate the odds that a specific case appears in one group over another. For example, you have fifty people and use a random process (such as the toss of a balanced coin) to place some in one (the coin that was always heads) or another group (the coin indicates tails). This way all participants have an equal chance of ending up in one or the other group.

The great thing about a random process is that over many separate random occurrences, very predictable things happen. Although the process is entirely due to chance and it is impossible to predict a specific outcome at a specific time, we can make highly accurate predictions when looking over many situations.

Random assignment is *unbiased* because our desires to confirm a hypothesis or a research participant's personal interests do not enter into the selection process. *Unbiased* does not mean the groups will be identical in each specific random assignment selection but is something close to this: We can determine the probability of selecting a case mathematically and, in the long run, across many separate selections, the average across all the groups will be identical.

Random sampling and random assignment are both processes for selecting cases for inclusion in a study. When we randomly assign, we sort a collection of cases into two or more groups using a random process. When we randomly sample, we select a smaller subset of cases from a far larger collection of cases (see Figure 1). We can both sample and

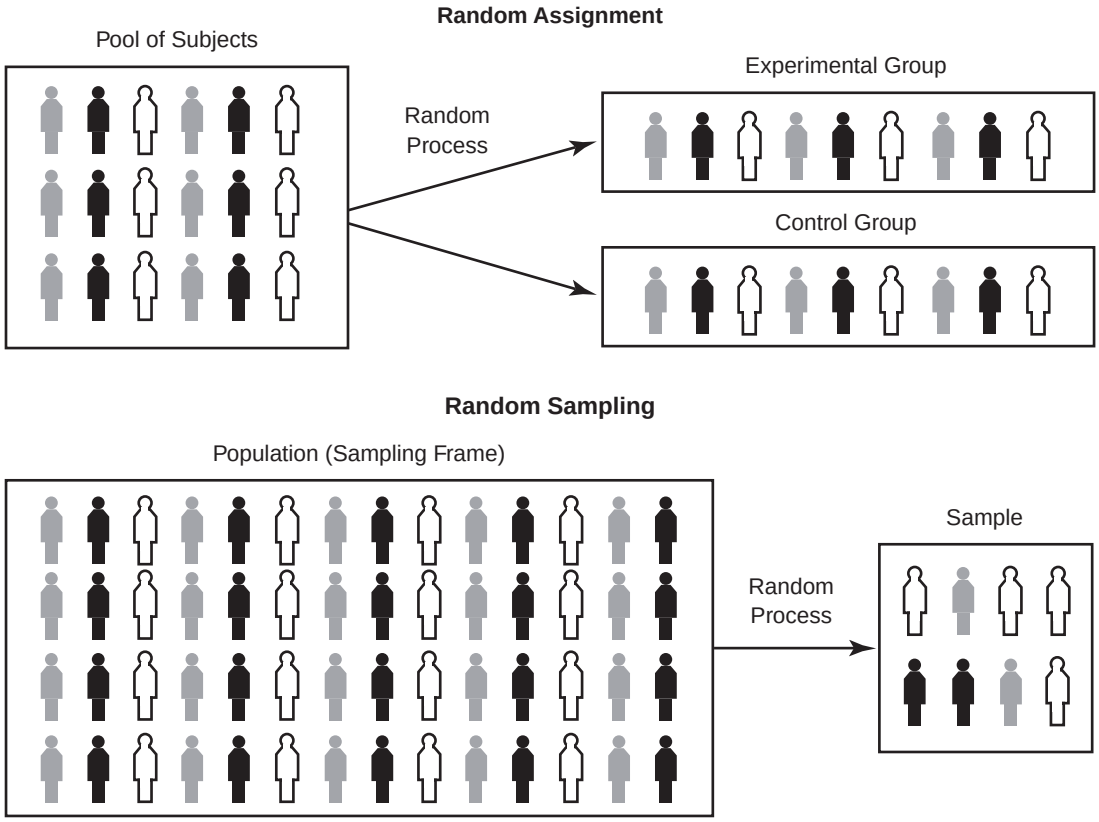


FIGURE 1 Random Assignment and Random Sampling

randomly assign. We 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).

How to Randomly Assign

Random assignment is simple in practice. We begin with a collection of cases (i.e., individuals, teams, companies, or whatever the unit of analysis is) and then divide the collection into two or more groups using a random process, such as asking people to count off, tossing a coin, or throwing dice. For example, we want to divide thirty-two people into two groups of sixteen. We could have each write his or her name on a standard size slip of paper, put all slips in a hat, mix the slips with our eyes closed, and then with eyes still closed, draw the first sixteen names for group 1 and the second sixteen for group 2. A specific situation can be unusual and the groups may differ. For example, it is possible although extremely unlikely that all cases with one characteristic will end up in one group. For example, we have thirty-two people with sixteen males and sixteen females, but all of the males end up in one group and all of the females in another. This is possible by random chance but extremely rare (see in Figure 2 on random assignment).

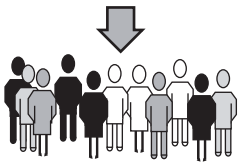
Matching versus Random Assignment

If the purpose of random assignment is to get two (or more) equivalent groups, you may ask whether

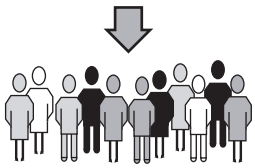
Step 1: Begin with a collection of subjects.

Step 2: Devise a method to randomize that is purely mechanical (e.g., flip a coin).

Step 3: Assign subjects with "Heads" to one group and "Tails" to the other group.



Control Group



Experimental Group

Note: Shading indicates various skin tones.

FIGURE 2 How to Randomly Assign

it would 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 gender. Matching is an alternative to random assignment, but it is an infrequently used one.

Matching presents a problem: What are the relevant characteristics on which to match, and can one locate exact matches? Individual cases differ in thousands of ways, and we cannot know which might be relevant. For example, we compare two groups of fifteen students. Group 1 has eight males, so we need eight males in group 2. Two males in group 1 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 Catholic. To match groups, do we have to find a tall Jewish male only child from a divorced home and a short Catholic male only child from an intact home? The tall, slender, Jewish male child is only 22 years old, and he is a premed major. The short, heavy Catholic male is 20 years old and is an accounting major. Do we also need to match the age and career aspirations of the two males? True matching soon becomes an impossible task.

Subjects A traditional name for participants in experimental research.

Treatment The independent variable in experimental research.

EXPERIMENTAL DESIGN LOGIC

The Language of Experiments

In experimental research, many studies call the participants **subjects**, although in recent years, *research participant* has been more commonly used**.**

Parts of the Experiment. Experiments have seven parts. Not all experiments have all of these parts, and some have all seven parts plus others.

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

In most experiments, we create a situation or enter into an ongoing situation and modify it. The **treatment** (or the stimulus or manipulation) is what we do. The term comes from medicine: a physician administers a treatment to patients; the physician

intervenes with a physical or psychological treatment to change it. The treatment is the independent variable or a combination of independent variables. In the study described in this chapter's opening box, Pager (2007) had two independent variables: one was a fixed characteristic (the tester's race) and the other was manipulated (a criminal conviction on a false résumé). In Niven's study (2002) (Example Box 1), the treatment was which of three news stories participants received to read while in an airport waiting area. In Transue's study (2007) (Example Box 2), the treatment was which of two questions about identity participants heard in a telephone survey.

At times, we go to great lengths to create treatments. While some may use reading different false records, reading different news stories, hearing different survey questions, or seeing different videos (see Example Box 4). Other treatments can be as complex, such as putting participants into situations with elaborate equipment, staged physical settings, or contrived social situations. See the Milgram and Zimbardo experiments in Example Box 6 later in this chapter). We want the treatment to have an impact and produce specific reactions, feelings, or behaviors (see the section on experimental realism later in this chapter).

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

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

We often divide participants 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 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, we decide on a topic, narrow it into a testable research problem or question, and then develop a hypothesis with variables. A crucial early step is to plan a specific experimental design (to be discussed). As we plan, we decide 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 participants will experience from beginning to end of the study. We often *pilot test* the experiment (i.e., conduct it as a "dry run").

The experiment begins after we locate volunteer participants and randomly assign them to groups. We give them precise, preplanned instructions. Next we may measure the dependent variable in a pretest before the treatment. We then expose one group only to the treatment (or a high level of it). Finally, we measure the dependent variable in a posttest. We also interview participants about the experiment before they leave. We record measures of the dependent variable and examine the results for each group to see whether the hypothesis is supported.

Control in Experiments. Control is crucial in experimental research.[6] We want to control all aspects of the experimental situation to isolate the effects of the treatment. By controlling confounding

Pretest An examination that measures the dependent variable of an experiment prior to the treatment.

Posttest An examination that measures the dependent variable of an experiment after the treatment.

Experimental group The participants who receive the treatment in experimental research.

Control group The participants who do not receive the treatment in experimental research.

variables, we eliminate alternative explanations that could undermine our attempts to establish causality.

We sometimes use deception to control the experimental setting (see the section A Word on Ethics later in this chapter). **Deception** occurs when we intentionally mislead research participants through written or verbal instructions, the actions of others, or aspects of the setting. Using deception may involve the use of a **confederate**—someone who pretends to be another research participant or bystander but who actually works for the researcher and deliberately misleads participants. Milgram's experiment used confederates as did the study described in Example Box 6 later in this chapter about disabled co-workers.

The purpose of deception is to control what the participants see and hear and what they believe is occurring. This usually means creating a **cover story**, a false explanation of the study's purpose that we tell participants to mislead them about its true purpose. The cover story helps satisfy curiosity but reduces demand characteristics (see later in this chapter). Many studies use a cover story (see studies in Example Boxes 1, 4, 6, and 7).

Types of Design

We combine parts of an experiment (e.g., pretests, control groups) into an **experimental design**. Some designs lack pretests, some do not have control groups, and others have many experimental groups. We have given widely used standard designs names. It is important to learn the standard design for two reasons. First, when reading research reports, researchers may name a standard design instead of describing it. Second, the standard designs illustrate common ways to combine design parts. We can use them for experiments we conduct or create variations.

Deception A lie by an experimenter to participants about the true nature of an experiment or the creation of a false impression through his or her actions or the setting.

Confederate A person working for the experimenter who acts as another participant or in a role in front of participants to deceive them with an experiment's cover story.

Cover story A type of deception in which the experimenter tells a false story to participants so they will act as wanted and do not know the true hypothesis.

Experimental design The planning and arranging of the parts of an experiment.

Classical experimental design An experimental design that has random assignment, a control group, an experimental group, and a pretest and posttest for each group.

We illustrate the various designs with a simple example. Let us say that you want to learn whether waitstaff (waiters and waitresses) receive more in tips if they first introduce themselves by first name and return 8 to 10 minutes after delivering the food to ask, "Is everything fine?" The dependent variable is the size of the tip received. Your 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. You give forty newly hired waitstaff an identical 2-hour training session and instruct the members 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. You next randomly divide the servers into two equal groups of twenty and send each group to one of the two restaurants to begin employment. You record the amount in tips for all participants for one month (pretest score). Next, you "retrain" the twenty participants at restaurant 1 (experimental group). You instruct 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). You remind the group at restaurant 2 (control group) to continue without an introduction or checking during the meal. Over the second month, you record the amount of tips for both groups (posttest score).

Preexperimental Designs. Some designs lack random assignment and are compromises or shortcuts.

We use these **preexperimental designs** in situations in which it is difficult to use the classical design. The designs have weaknesses that make inferring a causal relationship 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. You take a group of forty newly hired waitstaff and give all a 2-hour training session in which you instruct them 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 participants begin employment, and you record the amount in tips for all 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. You take a group of forty newly hired wait staff and give all a 2-hour training session. You instruct the staff members 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 you record the amount in tips for all for one month (pretest score). Next, you "retrain" all 40 participants and instruct 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). Over the second month, you record the amount of tips for both groups (posttest score).

This is an improvement over the one-shot case study because you measure the dependent variable before and after the treatment. But it lacks a control group. We 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,* a **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. You give forty newly hired waitstaff an identical 2-hour training session and instruct all to follow a script in which servers are not to introduce themselves by first name and but to return during the meal to check on the customers. They can choose one of the two restaurants at which to work, as long as each restaurant has twenty people. All begin employment. After one month, you "retrain" the twenty participants at restaurant 1 (experimental group) and instruct 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 "retrained" to continue without an introduction or checking during the meal. Over the second month, you record the amount of tips for both groups (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 us test for causal relationships in situations in which the classical design is difficult or inappropriate. We call them *quasi* because they

Preexperimental designs Experimental plans that lack random assignment or use shortcuts and are much weaker than the classical experimental design; are substituted in situations in which an experimenter cannot use all of the features of a classical experimental design but the design has weaker internal validity.

One-shot case-study design An experimental plan with only an experimental group and a posttest but no pretest.

Static group comparison design An experimental plan with two groups, no random assignment, and only a posttest.

Quasi-experimental designs Plans that are stronger than preexperimental ones; variations on the classical experimental design used in special situations or when an experimenter has limited control over the independent variable.

TABLE 1 A Comparison of the Classical Experimental Design

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

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 1).

Two-Group Posttest-Only Design. This design is identical to the static group comparison with one exception: You randomly assign. It has all parts of the classical design except a pretest. Random assignment reduces the chance that the groups differed before the treatment, but without a pretest, you cannot be as certain that the groups began the study at the same level on the dependent variable.

In a study using a two-group posttest-only design with random assignment, Rind and Strohmetz (1999) examined restaurant tips. The treatment involved messages about an upcoming special written on the back of customers' checks. The participants were eighty-one 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 the tip. The researchers gave a server with two years' experience a randomly shuffled stack of cards. One-half said No Message and one-half 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 upcoming 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 design An experimental plan in which the dependent variable is measured periodically across many time points and the treatment occurs in the midst of such measures, often only once.

Equivalent time-series design An experimental plan with several repeated pretests, posttests, and treatments for one group often over a period of time.

Interrupted Time Series. In an **interrupted time-series design**, you measure the dependent variable on one group over time using many multiple dependent variable measures before (prettests) and after a treatment (posttests).

Equivalent Time Series. An **equivalent time-series design** is a one-group design similar to the interrupted time series design. It extends over a time period, but instead of a single treatment, the equivalent time series design has the same treatment multiple times. Like the interrupted time series design, we measure the dependent variable several times before and after the treatments. The study on alcohol sales and suicide rates (Example Box 3,

EXAMPLE BOX 3

Interrupted Time Series, Alcohol Sales, and Suicide Rates

Governments face strong pressures by economic interests to modify laws to allow them to collect increased profits from alcohol sales. In most of western Canada, a public monopoly controlled alcohol sales and distribution through most of the twentieth century. Proponents of privatization point to its economic benefits, including selling previously government-owned retail outlets and the sale of licenses to merchandise alcohol. Others point to the impact of privatization on consumption and health. Studies of privatization of sales of alcoholic beverages indicate that privatization greatly expands alcohol availability and consumption.

Alberta moved to privatize alcohol sales in three stages: the opening of privately owned wine stores in 1985, the opening of privately owned cold beer stores and sale of spirits and wine in hotels in the rural area in 1989–1990, and finally the privatization of all liquor stores in 1994. The number of alcohol outlets increased substantially, and consumption of spirits increased dramatically at a time when consumption was decreasing elsewhere in the country. Privatization in Alberta has been associated with an increase in criminal offenses, such as liquor store break-ins and less strict enforcement of underage purchase laws. Alberta also has some of the highest rates of drunk-driving fatalities in the country. Many past studies also showed a strong relationship between suicide rates and alcohol consumption.

Zalcman and Mann (2007) used a three-stage interrupted time-series design to examine the influence of Alberta's privatization of alcohol sales on suicide rates between 1976 and 1999. They considered whether suicide rates changed after each privatization phase. They also compared Alberta's suicide levels to those for the same years in Ontario where alcohol sales remained a government monopoly.

The researchers found that the 1985 privatization of wine retailers increased male and female suicide rates in Alberta by 51 percent for males and 35 percent for females. After the 1989–1990 privatization of spirits and wine a significant increase occurred in male and female suicide rates, estimated to be 17 percent and 52 percent, respectively. The 1994 privatization event significantly increased male suicide mortality rates, estimated at 19 percent, but not female suicide rates. Part of the increase was a short-term spurt but long-term suicide raises also rose. By tracing the rates both over time by comparing those in a "control group" or to those in Ontario, the authors provided evidence of the effect of alcohol privatization.

Interrupted Time Series, Alcohol Sales, and Suicide Rates) illustrated equivalent time series.

Latin Square Design. At times, we are interested in how several independent variables in different sequences or time orders affect a dependent variable. The **Latin square design** enables us to examine this type of situation. 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, and then the LL system. In another class, using a compass comes first, then map reading, and then using 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 to each class 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 could see whether presenting units in one sequence or another resulted in improved learning.

Solomon Four-Group Design. We believe that the pretest measure may have an influence on the treatment or dependent variable. A pretest can sometimes sensitize participants to the treatment or improve their performance on the posttest (see the

Latin square design An experimental plan to examine whether the order or sequence in which participants receive versions of the treatment has an effect.

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 participants to one of four groups. For example, a mental health worker wants to find out 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 these groups 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 had an effect on the dependent variable.

Factorial Designs. Sometimes we are curious about the simultaneous effects of two or more independent variables. A **factorial design** uses two or more independent variables in combination. We look at each combination of the categories in variables (sometimes called *factors*). When each variable contains several categories, the number of combinations grows quickly. In this type of design, the treatment is not each independent variable; rather, it is each combination of the variable categories. 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 $2 \times 3 \times 3$ design means that there are three independent variables, one with two categories and two with three categories each.

For example, Krysan and associates (2009) wanted to study neighborhood preferences, but it was difficult to examine both racial and social class features of a neighborhood at the same time, so they used a factorial design (see Example Box 4, Factorial Experiment on Neighborhood Preference). The three independent variables of their study were participant race (two categories, Black or White), neighborhood composition (three types, all White, all Black, racially mixed), and social class (5 levels). The dependent variable was the desirability of a neighborhood based on a rating of 1 to 7. They had a $2 \times 3 \times 5$ factorial design. (The authors also asked participants about the strength of their identity with their own racial group.)

In a factorial design, treatments can have two types of effects on the dependent variable: main effects and interaction effects. Only *main effects* are present in one-factor or single-treatment designs. In other words, we simply examine the impact of the treatment on the dependent variable. In a factorial design, specific combinations of independent variable categories can have an effect beyond a single factor effect. We call them **interaction effects** because the categories in a combination interact to produce an effect beyond that of each variable alone. Interaction effects are of special interest because they suggest that not only an independent variable has an impact but also specific combinations have unique effects, or variables only have an impact under certain conditions.

Mueller-Johnson and Dhami (2010) (see Example Box 5, Mock Jury and Interaction Effects by Age and Crime) created a mock jury with a

Solomon four-group design An experimental plan in which participants are randomly assigned to two control groups and two experimental groups; only one experimental group and one control group receive a pretest; all four groups receive a posttest.

Factorial design An experimental plan that considers the impact of several independent variables simultaneously.

Interaction effect A result of two independent variables operating simultaneously and in combination on a dependent variable; is larger than a result that occurs from the sum of each independent variable working separately.

EXAMPLE BOX 4

Factorial Experiment on Neighborhood Preference

Krysan and associates (2009) created an experiment to study neighborhood preferences among Blacks and White adults in the United States. Past studies had looked at this issue; however, examining both racial and social class factors at the same time was very difficult, and telling whether people preferred a neighborhood for its social class or its racial features was not possible. The authors said, "At the core of our analysis are two research questions: (1) Are neighborhood preferences color blind or race conscious? (2) If preferences are race conscious, do they reflect a desire to be in a neighborhood with one's 'own kind' or to avoid being in a neighborhood with another racial group?" (p. 529). In 2004–2005, the authors selected more than 700 participants in the Detroit region and nearly 800 in the Chicago metropolitan area. To disentangle the class and race effects in neighborhoods, the authors showed participants videotaped neighborhoods that varied by social class and racial mix. They created thirteen videos in total. The neighborhoods varied by five social class levels and three racial mix levels.

> *We selected different neighborhoods to convey the different social class levels, relying on this assumption that respondents infer social class based on features such as home and property size, upkeep of the houses, and other cues gleaned from observation. Each of the different neighborhoods had, in turn, three variants in terms of the race of the individuals shown: (1) all residents are white; (2) all residents are black; (3) three residents are white and two residents are black. (p. 537)*

One video was a control without people. In each other video, five people (actors) appeared as residents engaged in ordinary activities. They noted (p. 537),

> *In each neighborhood, there was one scene in which three individuals were shown together talking in the driveway, in the front yard, at the mailbox, or surrounding a car that was being repaired. Residents wore short-sleeved shirts and no hats to increase the likelihood that the respondents could detect their racial/ethnic identity. Residents within each neighborhood social class level were matched on approximate age, gender, and style of dress.*

As a manipulation check, the authors showed videos to a small group of other participants prior to the actual study to verify that people saw the class and race composition of neighborhoods as intended. After viewing videos, the authors asked participants to rate each neighborhood on a seven-point Likert scale from very desirable to very undesirable. They said (p. 539), "Our dependent variables are the desirability ratings of the four neighborhoods, and thus our unit of analysis is the video. Given that each respondent saw and rated the same baseline video—an upper-working-class neighborhood with no residents—we include the ratings of this neighborhood as a respondent-level control." The authors used a factorial design with three independent variables: research participant race, neighborhood social class, and neighborhood racial mix. The authors randomly assigned participants to view different racial compositions in the same neighborhoods. Among their many findings, the authors note (p. 538), "Our fundamental conclusion is that race, per se, shapes how whites and, to a lesser extent, blacks view residential space. Residential preferences are not simply a reaction to class-based features of a neighborhood; they are shaped by the race of the people who live there."

trial-like situation and participants as a jury. The researchers presented various combinations of characteristics of offenders to see their impact on sentencing decisions (see Figure 3). The authors varied the age, health, offense severity, and prior convictions of an offender to create a $2 \times 2 \times 2 \times 2$ factorial design. They found main effects for severity of crime, age, and prior conviction. People committing more severe crimes, younger offenders, and those with prior convictions received longer sentences than people committing less serious crime, older offenders, and those with no prior convictions. They also found a few interaction effects; one was age and severity of crime for those with a past conviction.

EXAMPLE BOX 5

Mock Jury and Interaction Effects by Age and Crime

Mueller-Johnson and Dhami (2010) created a mock jury. They formed a trial-like situation and had participants form a jury. The authors presented various combinations of characteristics of offenders to see how they impacted jury sentencing decisions. Sentencing was length of prison term. Their jurors were forty-seven students (thirty-six women and eleven men) from an English university. The authors varied the age, health, offense severity, and prior convictions of an offender to create a $2 \times 2 \times 2 \times 2$ factorial design. In past experiments, they had found main effects for health, prior convictions, and severity of offense. People in poor health received shorter sentences, and older (66- to 72-year-old) received shorter sentences than younger (21- to 26-year-old) offenders regardless of the number of prior convictions. Younger offenders with prior convictions and more severe offences received longer sentences. In the current study, they investigated child sex offenders. Prior offense was either no prior conviction or one for sexual contact with a child 4 years earlier, and offense severity was either once touching a 7-year-old girl's genitals over her clothing or touching naked genitalia ten times over the course of a year. The participants usually decided on a sentence in 15 minutes. The authors found interesting interaction effects among age, offense severity, and previous convictions. For those with a prior conviction, older offenders received a longer sentence than younger offenders with less serious offenses, but shorter sentences if the offense was more serious. In other words, the combination of a prior conviction and less serious offense for older offenders resulted in a longer sentence. This is consistent with the "dirty-old-man" stereotype.

FIGURE 3 Sentence in Mock Jury Trial for Sex Offenders with One Prior Conviction

Design Notation

We can design experiments in many ways. **Design notation** is a shorthand system for symbolizing the parts of experimental design.[7] It expresses a complex, paragraph-long description of the parts of an experiment in five or six symbols arranged in two lines. Once you learn design notation, you will find it easier to think about and compare designs. Design notation 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. We arrange symbols in rows with each row representing a group of participants. 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 2 gives the notation for many standard experimental designs.

Design notation A symbol system used to show parts of an experiment and to make diagrams of them.

Internal validity The ability of experimenters to strengthen the logical rigor of a causal explanation by eliminating potential alternative explanations for an association between the treatment and dependent variable through an experimental design.

INTERNAL AND EXTERNAL VALIDITY

The Logic of Internal Validity

Internal validity occurs when the independent variable, and nothing else, influences the dependent

TABLE 2 Summary of Experiment Designs with Notation

NAME OF DESIGN	DESIGN NOTATION
Classical experimental design	R O X O R O O
Preexperimental designs	
One-shot case study	X O
One-group pretest/posttest	O X O
Static group comparison	X O O
Quasi-experimental designs	
Two-group posttest only	R X O O
Interrupted time series	O O O O X O O O
Equivalent time series	O X O X O X O X O
Latin square designs	O X_a O X_b O X_c O O X_b O X_a O X_c O R O X_c O X_b O X_a O O X_a O X_c O X_b O O X_b O X_c O X_a O O X_c O X_a O X_b O
Solomon four-group design	O X O R O O X O O
Factorial designs	X_1 Z_1 O R X_1 Z_2 O X_2 Z_1 O X_2 Z_2 O

Note: Subscripts with letters indicate different treatment variables. Subscripts with numbers indicate different categories of the same treatment variable, such as male or female for gender.

variable. Anything other than the independent variable influencing the dependent variable threatens internal validity. These are confounding variables; they confound the logic of an experiment to exclude everything except the relationship between the variables in your hypothesis. They threaten your ability to say that the treatment was the true causal factor that produced a change in the dependent variable. You may also hear them called **artifacts**. This is

Artifact An object in experimental research studies; refers to the type of confounding variable that is not part of the hypothesis but affects the experiment's operation or outcome. In field research studies, it refers to physical objects that humans created that have cultural significance; specifically, objects that members use or to which they attach meaning that we study to learn more about a cultural setting or its members.

because the unwanted or confounding variables do not come from the natural relationship you are examining but are due to the particular experimental arrangement. An artifact appears by accident because during preparation of the study, you unintentionally introduce something that alters things. For example, you clean a room before participants arrive for an experiment on the emotional effects of going without sleep, but the cleaning solution you used to wipe down tables and chairs causes irritability in many people. Your results show increased irritability among people who had little sleep. However, it is not because of sleep loss but an unintended side effect of your cleaning solution. You want to rule out artifacts and confounding variables—everything that could possibly affect the dependent variable other than the treatment. You rule out artifacts and confounding variables by controlling experimental conditions and by using experimental designs. Next we examine major threats to internal validity.

Threats to Internal Validity

The following are 12 threats to internal validity.[8]

1. *Selection bias.* **Selection bias** can arise when an experiment has more than one group of participants. You want to compare the groups, but they differ or do not form equivalent groups. This is a problem in designs without random assignment. For example, you design a two-group experiment on aggressiveness. If you do not use randomization or randomization is not effective, the treatment group could by chance differ. You may have sixty research participants who are active in various campus activities. By chance, many of your volunteers for the experimental group have participated in football, rugby, hockey, and wrestling whereas volunteers in your control group are musicians, chess club members, ballet dancers, and painters. Another example of selection bias is an experiment on the ability of people to dodge heavy traffic. Selection bias would occur if participants assigned to one group are from rural areas with little traffic experience and those in the other grew up in large cities and have traffic experience. You can often detect selection bias by comparing pretest scores. If you see no group differences in the pretest scores, selection bias is probably not a problem.

2. *History.* **History effect** is the result of 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. For example, halfway through a two-week experiment to evaluate feelings about pet dogs, a fire at a nearby dog kennel kills and injures many puppies with news reports showing injured animals and many local people crying over the incident.

3. *Maturation.* A **maturation effect** is a result of a threat that a biological, psychological, or emotional process within participants other than the treatment occurs during the experiment and influences the dependent variable. The time period for maturation effects to occur can be hours, months, or years depending on the dependent variable and study design. For example, during a daylong eight-hour experiment on reasoning ability, participants become bored and sleepy and, as a result, their scores are 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 maturational 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 to determine whether maturation or history effects are present because both experimental and control groups will show similar changes over time.

Selection bias A preconception that threatens internal validity when groups in an experiment are not equivalent at the beginning of the experiment with regard to the dependent variable.

History effect Result that presents a threat to internal validity because of something that occurs and affects the dependent variable during an experiment; is unplanned and outside the control of the experimenter.

Maturation effect A result that is a threat to internal validity in experiments because of natural processes of growth, boredom, and so on that occur during the experiment and affect the dependent variable.

4. *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 to detect testing effects. For example, you pretest to determine how much participants know about geology and geography. Your treatment is a series of videos about geology and geography viewed over 2 days. If participants 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, you cannot say that the treatment alone has affected the dependent variable. The dependent variable was influenced by both memory of the pretest and the treatment.

5. *Instrumentation.* This threat is related to stability 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 is a treatment to show a video, but the video equipment failed to work for some participants.

6. *Experimental mortality.* When some research participants do not continue throughout the entire experiment, **experimental mortality**, or attrition, arises. Although the word *mortality* means death, it does not necessarily mean that they have died. If many participants leave partway through an experiment, we cannot know whether the results would have been different had they stayed. For example, you begin a weight-loss experiment with sixty people. At the end of the program, forty remain, each of whom lost 5 pounds with no side effects. The twenty who left could have differed from the thirty who stayed, changing the results. Perhaps the program was effective for those who left, and they withdrew after losing 25 pounds. Or perhaps the program made them sick and forced them to quit, or they saw no improvement and dropped out. We need to notice and report the number of participants at all stages of an experiment to detect this threat to internal validity.

7. *Statistical regression effect.* This 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 in which **statistical regression effect** occurs is when participants are unusual with regard to the dependent variable. Because they are unusual, they do not respond further in one direction. For example, you want to see whether playing violent video games makes people more aggressive. Your participants are a group of convicts from a high-security prison. You give them a pretest, have them play 60 hours of extremely violent video games, and then administer a posttest. To your surprise, there is no change. It could be that the convicts started as extremely aggressive so your treatment could not make them any more aggressive. By random chance alone, some may even appear to be less aggressive when measured in the posttest.[9]

A second statistical regression effect situation involves a problem with the measurement instrument. If your measure is such that most people 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, you give eighty participants a simple math test, and seventy-seven get perfect scores. You give a treatment to improve math scores. Because so many already had perfect scores, random errors could reduce the group average because the seventy-seven who got perfect scores can move in only one direction—to get an answer wrong, and only three could improve. As a result, the group average may appear lower in the posttest due to chance alone. You need to monitor the range of scores to detect statistical regression.

Testing effect A result that threatens internal validity because the very process of measuring in the pretest can have an impact on the dependent variable.

Experimental mortality Threat to internal validity because participants fail to participate through the entire experiment.

Statistical regression effect A threat to internal validity from measurement instruments providing extreme values and a tendency for random errors to move extreme results toward the average.

8. *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. You can avoid this by isolating groups or having them promise not to reveal anything to other participants. For example, you have eighty participants in a daylong experiment on ways to memorize words. The treatment group is taught a simple method, but the control group is told to use any technique the members want to use. During a break, participants in the treatment group tell those in the control group about the new method. After the break, control group particpants start using it too. You might ask about possible diffusion in a post-experiment interview with participants to reduce this threat.

9. *Compensatory behavior.* In experiments that provide something of value to one group of participants but not to another and the difference becomes known, **compensatory behavior** is said to occur. The inequality between groups may create a desire to reduce differences, competitive rivalry between groups, or resentful demoralization. Such behavior can affect the dependent variable in addition to the treatment. For example, students in one school receive a treatment of longer lunch breaks to produce gains in learning, but students in another school have a regular lunchtime. Once the inequality is known, stundents in the control group (school without long lunch breaks) work extra hard to learn and to overcome the inequality. Alternatively, the control group students could become demoralized by the unequal treatment and put less effort into learning. It is difficult to detect this threat unless you obtain outside information (see the discussion of diffusion of treatment).

10. *Experimenter expectancy.* An experimenter's behavior might threaten internal validity if the experimentor indirectly communicates a desired outcome.[10] This is called **experimenter expectancy**. Because of a strong belief in the hypothesis, even the honest experimenter might unintentionally communicate desired findings. For example, you study participants' reactions to people with disabilities. You deeply believe that females are more sensitive to those with disabilities than males are. Through eye contact, tone of voice, pauses, and other nonverbal communication, you might unconsciously encourage female research participants to report positive feelings toward those with disabilities; your nonverbal behavior is the opposite for male participants.

The **double-blind experiment** is a design intended to control experimenter expectancy. In this experiment, the only people who have direct contact with participants do not know the details of the hypothesis or the treatment. It is *double* blind because both the participants and those in contact with them are blind to details of the experiment (see Figure 4). For example, you want to see whether a new drug is effective. Using pills of three colors—green, yellow, and pink—you put the new drug in the yellow pill, an old drug in the pink one, and make the green pill a *placebo* (i.e., an empty or nonactive treatment). Assistants who give the pills and record the effects do not know which color pill contains the new drug. They just administer the pills and record results by color of pill. Only you know which color pill contains the drug and examine the results, but you have no contact with the research participants. The double-blind design is nearly mandatory in medical research because experimenter expectancy effects are well recognized.

11. *Demand characteristics.* A threat to internal validity related to reactivity (discussed in next

Diffusion of treatment The spread of a threat to internal validity that occurs when the treatment "spills over" from the experimental group and control group participants modify their behavior because they learn of the treatment.

Compensatory behavior Conduct that is a threat to internal validity when participants in the control group modify their behavior to make up for not getting the treatment.

Experimenter expectancy A type of reactivity that occurs because the experimenter indirectly makes participants aware of the hypothesis or desired results.

Double-blind experiment A type of experimental research in which neither the participants nor the person who directly deals with them for the experimenter knows the specifics of the experiment.

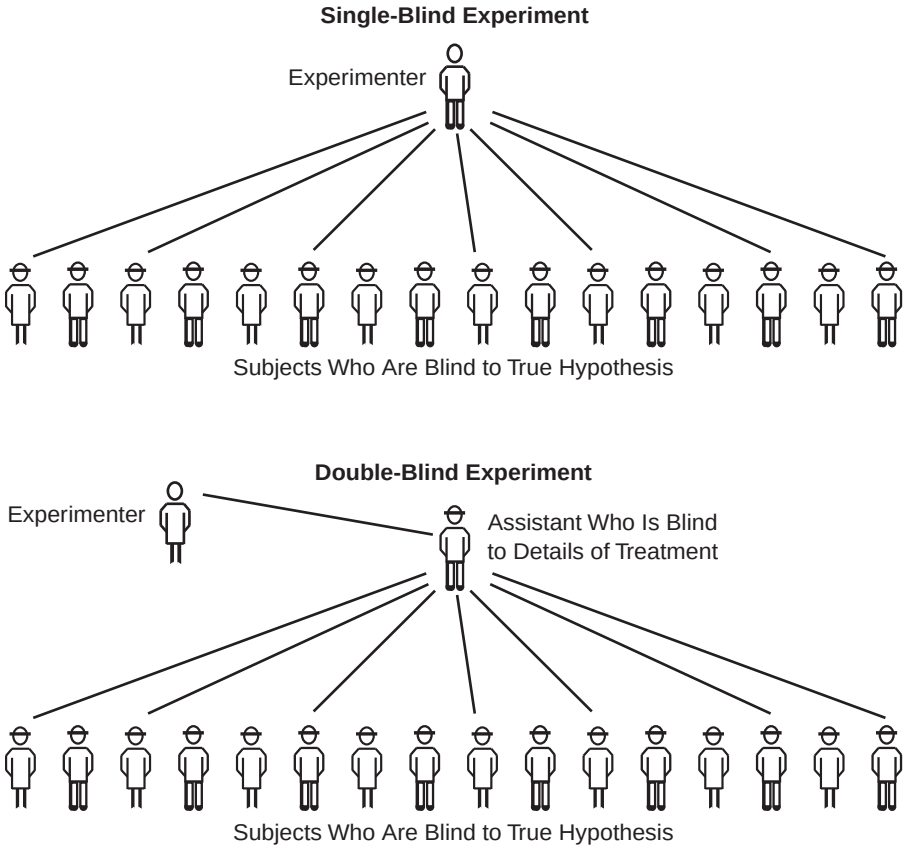


FIGURE 4 Double-Blind Experiment: An Illustration of Single-Blind, or Ordinary, and Double-Blind Experiments

section of this chapter) is called a **demand characteristic**. It occurs when research participants pick up clues about the hypothesis or an experiment's purpose and then modify their behavior to what they think the research demands of them (i.e., support the hypothesis). Participants often do this to please the researcher, which is why we often use mild deception in the form of cover stories.

12. *Placebo effect.* The last type of threat to internal validity is the **placebo effect**. A *placebo* is an empty or nonactive treatment, such as a sugar pill in medical research. It occurs when you give some participants a placebo but they respond as if they have received the real treatment. For example, you create an experiment on stopping smoking for heavy smokers. You give some participants a pill with an antinicotine drug to reduce their nicotine dependence and others a placebo (empty pill). If participants who received the placebo also stop smoking, then merely participating in the experiment and taking something that they believed would help them quit smoking had an effect. The belief in the placebo alone may have affected the dependent variable (see Table 3 for a summary of internal validity threats).

Demand characteristic A type of reactivity in which participants in experimental research pick up clues about the hypothesis and alter their behavior accordingly.

Placebo effect A result that occurs when participants do not receive the real treatment but receive a nonactive or imitation treatment but respond as though they have received the real treatment.

TABLE 3 Internal Validity and External Validity Issues

INTERNAL VALIDITY	EXTERNAL VALIDITY
Selection bias	Population generalization
History effect	Naturalistic generalization
Testing effect	Theoretical generalization
Maturation effect	Mundane realism
Instrumentation	Experimental realism
Experimental mortality	Hawthorne effect
Statistical regression effect	
Diffusion of treatment	
Compensatory behavior	
Experimenter expectancy	
Demand characteristics	
Placebo effect	

Experimenters often undertake manipulation checks to increase internal validity. A **manipulation check** is a process to verify theoretically salient variables (e.g., independent, dependent, and intervening variables in hypotheses). Its purpose is to verify measurement validity (e.g., variables truly measure the theoretical concepts) of whether the conditions of the experiment had the intended effects, or the degree of its experimental realism (experimental realism is discussed later in this chapter). We have manipulation checks to make certain that the variables and conditions in our experiment operate as we intended and help us rule out possible threats to internal validity.

We check "manipulations" (our measures and interventions in an experimental situation) with pretests, pilot tests, and experimental debriefing. We might create a pretest of certain experimental conditions. For example, you have a confederate act as if he or she is disabled and have preliminary research participants observe the confederate. As a check, you ask whether the participants believed the confederate was truly disabled or just acting. In the study on neighborhood preference (see Example Box 3), the researchers showed videos of neighborhoods to a small number of people before using the videos in the study. This was done to verify that people recognized the racial mix and neighborhood's social class as the researchers intended. If you plan to provide participants with written or oral instructions in an experiment, you might pretest them with a few preliminary participants. You can inquire about the clarity of the instructions and whether the participants understood them as you intended.

A "dry-run" or pilot test of the entire experimental procedure can be a manipulation check. During and after the pilot test, you look for potential flaws, mishaps, or misunderstandings. You ask whether all parts of the experimental situation went smoothly and had their intended effects on participants. You may check to see whether participants paid attention and accepted the "cover story" if you used deception.

Experimental debriefing after a pilot test or the actual experiment can be a manipulation check. To conduct an experimental debriefing (unlike ethical debriefing that emphasizes removing a lie or deception), you interview participants about details of the experiment. You want to learn what they thought was happening, whether they felt fully engaged and took the situation seriously, and whether they felt any confusion, anxiety, or discomfort. You may discuss compensatory behavior and demand characteristics or diffusion of treatment in such interviews. At times, experimenters drop a participant from study data if they learn that the participant misunderstood a critical aspect of the experiment, saw through the cover story of deception, or modified responses because of demand characteristics (also see discussions on reactivity later in this chapter). For example, an experimenter may drop data of a participant who revealed that she or he did not accept the deception cover study but believed the study was about reactions toward disabled people (which it was) and responded based on that belief (i.e., showed demand characteristics) (see Example Box 6, Who Helps a Co-Worker Who Is Disabled?).

Manipulation check A separate measure of independent or dependent variables to verify their measurement validity and/or experimental realism.

EXAMPLE BOX 6

Who Helps a Co-Worker Who Is Disabled?

Miller and Werner (2007) conducted a laboratory experiment on helping behavior with two treatment conditions and a control group. The authors wanted to learn what types of people would be likely to assist a co-worker who is disabled. Past studies have found a positive relationship between personality traits and attitudes toward persons with disabilities. The researchers measured three personality traits: equity preference, feminine traits, and impression management. *Equity preference* comes from the idea that each person must do an amount of work for a reward. Some people are more benevolent (i.e., people who try harder should get equal rewards even if they produce less) and some feel more entitled (i.e., no one should receive a bigger reward if they do less). Traditional *feminine traits* are to be kind, helpful, and understanding. *Impression management* is a conscious representation of oneself to others. Those who score high on impression management act consciously to display an intended image in a public setting. The authors had more than 500 students in three sections of an undergraduate business management course complete a survey that measured personality traits. From these, the authors selected 133 volunteers for the experiment. They manipulated three levels of disability, their key independent variable: no disability, a mental disability, and a physical disability. They also did a *manipulation check* by asking a separate group of eighty-four participants to read descriptions of various people and rate the descriptions of the persons as being physically disabled, mentally disabled, or not disabled. The authors reported (p. 2668) that

> *to reduce the confounding of variables, the same confederate was used in each session of the experiment. This confederate was a male graduate student in a nonbusiness doctoral program at the university. The same confederate was used so that there was no variability on race, physical attractiveness, personality, and other characteristics that might have elicited differences in responses from participants. The confederate was a White student with a slight build who was 25 years old.*

At the beginning, each participant and the confederate prepared and read an autobiography. In the physical disability condition, the confederate was in a wheelchair and had an autobiography that included a past automobile accident that had left him wheelchair bound. In the mental disability condition, the confederate displayed difficulty with the autobiography and reported that he was in an automobile accident that had left him with a brain injury and short-term memory difficulties. Next, the participants and the confederate were to complete a complex paper-folding and envelope-stuffing task that required some physical movement and mental counting. Each person was told she or he would be paid for completing the task and had to finish it in exactly 5 minutes. The task required rapid work but was fairly easy to complete in the allotted time. In the physical disability condition, the wheelchair-bound confederate had difficulty moving to complete the task. In the mental disability condition, the confederate showed great difficulty in performing the mental calculations needed to complete the task in time. In the no-disability condition, the confederate just moved slowly. For all three conditions, it was clear to participants that the confederate could not complete the task on time. The dependent variable was whether any participants assisted the confederate. The researchers videotaped sessions and a trained, independent observer scored the amount of assistance participants gave to the confederate. Results showed that equity preference and impression management but not feminine traits had an effect. People high on benevolent equity preference and impression management helped more. The physically disabled condition received more help than the mentally disabled condition, and both disabled conditions received more than the nondisabled condition. In a debriefing interview after the experiment, researchers told participants the study's true purpose and asked what they thought the study was about. Researchers discarded data for five participants, "because they offered a comment that revealed that their ratings might have been biased. Examples of such comments include 'I thought that the disabled student was a decoy,' 'I think you wanted to see how we react to working with a disabled person. . . .'" (p. 2671).

External Validity and Field Experiments

Even if we eliminated all internal validity concerns, external validity would remain an issue. **External validity** is the effectiveness of generalizing experimental findings. If a study lacks external validity, the findings may hold true for only a specific experiment. Because we seek general theoretical knowledge in basic research and findings that relate to real-life problems in applied research, findings lacking external validity are nearly useless. However, in the widely cited article "In Defense of External Invalidity," Mook (1983) argued that generalizing from an experiment to natural, real-life settings is not a goal for many experiments. Instead, we may have other theoretical purposes (see later section on theoretical generalization).

The issue of external validity can be complex. Indeed, Thye (2007:81) says, "Perhaps the most misunderstood issue surrounding experiments is that of external validity." The reason is that external validity can involve several forms of generalization.[11] External validity addresses three questions about generalizing: Can we generalize from the specific collection of participants in an experiment to an entire population? Can we generalize from what occurs in a highly controlled and artificial experimental setting to most natural, "real-world" situations? Can we generalize from the empirical evidence of a specific experiment to an abstract theoretical model about relationships among variables? To address these questions, we can think of external validity as involving three forms of generalization that do not always overlap: populational, naturalistic, and theoretical (see Figure 5).

Populational Generalization. The key question for this form of external validity is whether we can accurately generalize from what we learn with a specific collection of people in one study to a universe or population of people/cases. To generalize the findings, we should specify the universe to which we wish to generalize and consider providing evidence to support such a generalization. For example, we conduct an experiment with one hundred undergraduate volunteers from one course in one university. To whom can we generalize these findings? To all undergraduate students in all courses at the same university during the same year, to all college students in the same country in the same decade, or to all humanity for all time? To improve the populational generalization form of external validity in an experiment, we would draw a random sample from a population and conduct the experiment on sampled participants.

Naturalistic generalization is what most people first think of when hearing the term *external validity*. The key question of naturalistic generalization is whether we can generalize accurately from what we learn in an artificially created, controlled laboratory-like setting to "real-life" natural settings. For naturalistic generalization, we need to consider two issues: mundane realism and reactivity.

Mundane realism asks whether an experiment or a situation is like the real world. For example, your study of learning has participants memorize four-letter nonsense syllables. Mundane realism would be stronger if you had them learn real-life factual information rather than nonsense syllables invented for an experiment alone.[12]

Reactivity is the effect of people responding because they are aware that they are in a study. Research participants might react differently in an experiment than in real life because they know someone is studying them. The **Hawthorne effect** is a specific kind of reactivity.[13] The name comes from a series of experiments by Elton Mayo at the Hawthorne, Illinois,

External validity The ability to generalize findings beyond a specific study.

Naturalistic generalization The ability to generalize accurately from what was learned in an artificially created controlled laboratory-like experimental settings to "real life" natural settings.

Mundane realism A type of external validity in which the experimental conditions appear to be real and very similar to settings or situations outside a lab setting.

Reactivity A result that occurs because of a general threat to external validity that arises because participants are aware that they are in an experiment and being studied.

Hawthorne effect A reactivity result named after a famous case in which participants responded to the fact that they were in an experiment more than to the treatment.

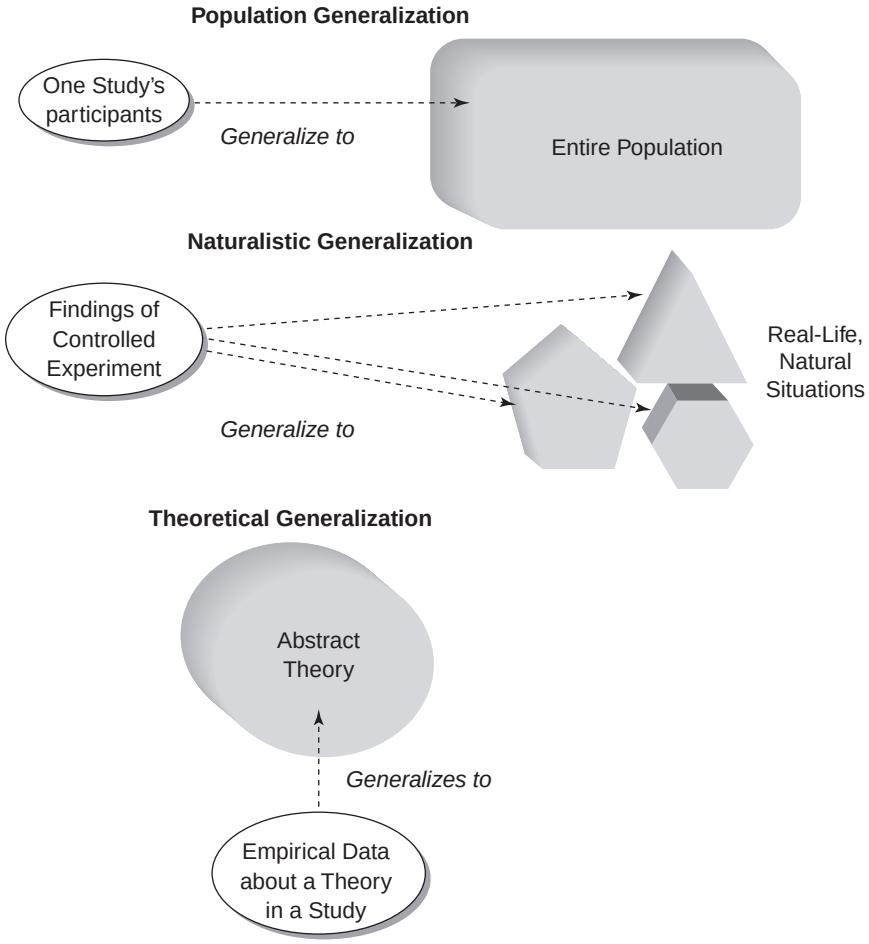


FIGURE 5 Three Types of External Validity Generalization

plant of Westinghouse Electric during the 1920s and 1930s. Researchers modified many aspects of working conditions (e.g., lighting, time for breaks) 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 the reported worker response had in fact occurred, but the name is still used for an effect that results from the attention of researchers.

For external validity concerns, the issue of reactivity is whether we can accurately generalize from activities that occur in a setting in which people are aware they are being studied to natural settings. Reactivity is most likely to occur in a highly controlled experiment in which the research participants know that an experimenter has created the conditions and is observing their behaviors or responses.

Let us say that you conduct an experiment in a college classroom or laboratory in which the participants know they are participating in a study. You ask the participants to engage in some artificially created tasks (e.g., assemble a puzzle) or create artificial status using deception (e.g., tell participants that a confederate working for you has a genius IQ score). After working on the task, you ask

participants to complete a questionnaire in which you have questions about their feelings regarding people with high IQ scores. To what settings in daily life might you generalize your study's findings? To all real-life workplace settings with people of varying intelligence levels, to all types of work tasks and all social statuses, or to all attitudes about other people naturally formed in daily life and retained in everyday thoughts, behavior, or conversations? To improve the naturalistic generalization form of external validity in an experiment, you would need to conduct a field experiment.

Theoretical generalization asks whether we can accurately generalize from the concepts and relations in an abstract theory that we wish to test to a set of measures and arrangement of activities in a specific experiment. This is probably the most difficult type of generalization because it includes several other ideas: experimental realism, measurement validity, and control-confounding variables (high internal validity). **Experimental realism** is the impact of an experimental treatment or setting on people; it occurs when participants are caught up in the experiment and are truly influenced by it. It is weak if they remain unaffected and the experiment has little impact on them.

Field Experiments. We conduct experiments under the controlled conditions of a laboratory and in real-life or field settings in which we have 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"—natural settings such as a subway car, a liquor store, or a public sidewalk. Participants 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.[14]

Some field experiments, such as those by Transue on racial identity and school taxes or Krysan and colleagues on neighborhood preference (see Example Boxes 2 and 3), involved gathering particpants and presenting them with realistic choices. Others are "natural experiments" in which experimental-like situations arise without total researcher control as with the Alberta privitization of alcohol sales (see Example Box 4). A related type of natural experiment in the field occurs when a researcher can take advantage of random assignment conditions of a key variable, as in the case of racial mixing of college roomates (see Example Box 7, A Field Experiment on College Roommates)

The amount of experimenter control is related to internal and external validity. Laboratory experiments tend to have higher internal validity but lower external validity. They are logically tighter and better controlled but less generalizable. Field experiments tend to have high external validity but low internal validity. They are more generalizable but less controlled. Quasi-experimental designs are more common. For example, in the experiment involving college roommates, the roommate situation was very real and lasted several months. The experiment had more external validity than putting people in a laboratory setting and asking them what they would do hypothetically.

Experimental realism External validity in which the experiment is made to feel realistic so that experimental events have a real impact on participants.

Laboratory experiment An experimental study in an artificial setting over which the experimenter has great control.

Field experiment A study that takes place in a natural setting.

PRACTICAL CONSIDERATIONS

Every research technique has "tricks of the trade" that are pragmatic strategies learned from experience. They account for the difference between the successful studies of an experienced researcher and the difficulties a novice researcher faces. Three are discussed here.

Planning and Pilot Tests

All social research requires planning. During the planning phase, we anticipate alternative explanations or threats to internal validity, develop a

EXAMPLE BOX 7

A Field Experiment on College Roommates

Contact hypothesis states that intimate, long-term contact with an out-group reduces prejudice. Shook and Fazio (2008) wanted "to assess the nature of interracial relationship and test the effect of intergroup contact" (p. 719). However, when we measure prejudice with self-report attitude measures, people often control prejudice reactions so they do not appear prejudicial even though they may harbor prejudicial attitudes. An indirect technique for measuring hidden or "automatic" racial prejudice measures the response time in seconds as a person sees visual images of people of different races matched with various adjectives (see Fazio et al., 1995). Speed of response indirectly measures racial prejudice because we respond more slowly as we try to hide true attitudes. To create a long-term field experiment, the authors took advantage of random assignment to college dormitory rooms and room shortage that prevented roommates from switching. The study had 136 White and 126 African American college freshmen. By random assignment, some had a same-race roommate, and others had a different race roommate. Roommate race was the independent variable. The authors had the students attend one session during the first two weeks and another during the last two weeks of the academic term. They asked students about several issues, including roommate satisfaction, activities with roommates, and social networks. The students also completed a questionnaire on racial attitudes and intergroup anxiety. In addition, the authors created a series of tasks asking students to respond to various images on a computer screen. After several such computer tasks to create a "cover story," a final task was to respond to images of faces matched with adjectives; one-half of the faces were African American and one-half White. This was the indirect measure of racial prejudice. Thus, the authors had multiple pretest and posttest measures of racial attitudes and interracial social interactions. As in past roommate studies, their results showed less social interaction and lower roommate satisfaction among the different race roommate pairs than same-race pairs. Over the academic term, satisfaction with same-race roommates declined slightly but for the different race roommates increased slightly. For roommates of a different race, intergroup anxiety declined and roommate social interactions increased over the three-month term. Both the direct and indirect measures of racial prejudice remained unchanged for same-race roommates. However, levels of prejudice declined significantly between the pretest and posttest for the students who had different race roommates, just as predicted by the contact hypothesis.

well-organized system for recording data, and pilot test any apparatus (e.g., computers, video cameras, tape recorders) that we will use. After the pilot test, we interview participants to uncover aspects of the experiment that need refinement.

Instructions to Participants

Most experiments involve giving instructions to participants to "set the stage." We must word instructions carefully and follow a prepared script so that all participants hear the exact same thing. This ensures reliability. The instructions are also important in creating a realistic cover story when deception is used. Aronson and Carlsmith (1968:46) noted, "One of the most common mistakes the novice experimenter makes is to present his instructions too briefly."

Postexperiment Interview

At the end of an experiment, we should interview participants for three reasons. First, if we used deception, we must ethically **debrief** the research participants (i.e., explain the true purpose of the experiment and answer any participants' questions). Second, we can learn what participants thought and how their definitions of the situation affected their behavior. Finally, we can explain the importance of not revealing the true nature of the experiment to other potential participants.

Debrief To gather information by talking with participants after an experiment to give a true explanation of the experiment if deception has been used or to learn their perceptions.

RESULTS OF EXPERIMENTAL RESEARCH: MAKING COMPARISONS

Comparison is critical to all research. By carefully examining the results of experimental research, we can learn about possible threats to internal validity and treatment effects on the dependent variable. In each study discussed in this chapter, the researchers carefully analyzed quantitative data to examine the effects of independent variables and considered potential internal validity concerns.

Here is an illustration of such comparisons (see Figure 6) based on the results of a series of five weight-loss experiments using the classical experimental design. In the example, the thirty research participants in the experimental group at Enrique's Slim Clinic lost an average of 50 pounds, whereas the thirty 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 eleven 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 eleven 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.

A WORD ON ETHICS

Ethical consideration is a significant issue in most experiments because they are often intrusive (i.e., interfere with ordinary activity). Experimental treatments may involve putting people in contrived social settings, asking them to engage in specific activities, or manipulating their feelings or behaviors. While doing this, we listen to what they say, observe their actions, and record responses. Ethical requirements limit the amount and type of allowable intrusion. We must never place research participants in physical danger, and we must take precautions when we put them in embarrassing or anxiety-inducing situations. It is essential to continuously monitor and control experimental events to ensure safe and ethical study.[15]

Sometimes we might use deception in social experiments by temporarily misleading participants. Such dishonesty might be acceptable but only if there is no other way to achieve a specific research goal. Even for a highly worthy goal, we only use deception with restrictions. The amount and type of deception cannot exceed the minimum needed for the specific purpose. In addition, we must always debrief research participants as soon as possible, telling them that they had been temporarily deceived and explaining the real situation to them.

CONCLUSION

This chapter discussed experimental research. In most experimental designs, we use random assignment to create two (or more) groups that we can treat as equivalent and hence compare. Experimental research provides precise and relatively unambiguous evidence for a causal relationship. It closely follows principles of a positivist approach to social science and produces quantitative results that we can analyze with statistics.

This chapter also examined how the parts of an experiment can be combined to produce different experimental designs. In addition to the classical experimental design, preexperimental and quasi-experimental designs and design notation were discussed.

Various threats to internal validity that are possible alternative explanations to the treatment were identified as were external validity and the ways that field experiments maximize naturalistic generalization in external validity.

The real strength of experimental research is its control and logical rigor in establishing evidence

FIGURE 6 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)	189 (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)

SYMBOLS FOR COMPARISON PURPOSES

	Pretest	*Posttest*
Experimental	A (A)	C (C)
Control group	B (B)	D (D)

COMPARISONS

	A–B	*C–D*	*A–C*	*B–D*	*(A)–(C)*	*(B)–(D)*
Enrique's	1	49	–50	0	–1	0
Susan's	1	48	–49	0	–11	0
Carl's	31	37	–8	–2	–1	0
Natalie's	2	1	–2	–3	–1	–1
Pauline's	1	1	–32	–32	0	–1

A–B	Do the two groups begin with the same weight? If not, selection bias may be possibly occurring.
C–D	Do the two groups end the same way? If not, the treatment may be ineffective, or there may be strong history, maturation, diffusion, or treatment effects.
A–C	Did the experimental group change? If not, treatment may be ineffective.
(A)–(C) and (B)–(D)	Did the number of participants in the experimental group or control group change? If a large drop occurs, experimental mortality may be a threat to internal validity.

INTERPRETATION

Enrique's:	No internal validity threats evident, shows effects of treatment
Susan's:	Experimental mortality threat likely problem
Carl's:	Selection bias likely problem
Natalie's:	No internal validity threat evident, shows no treatment effects
Pauline's:	History, maturation, diffusion of treatment threats are a likely problem

Note: Numbers are average number of pounds. Numbers in parentheses ( ) are number of participants per group. Random assignment is made to the experimental or control group.

for causality. In general, experiments tend to be easier to replicate, less expensive, and less time consuming than other research 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 a potential problem because many experiments rely on small nonrandom samples of college students.[16]

The chapter explained that careful examination and comparison of results can alert us to potential problems in research design. Finally, the chapter presented some practical and ethical considerations in experiments.

KEY TERMS

artifacts
classical experimental design
compensatory behavior
confederate
confounding variables
control group
cover story
debrief
deception
demand characteristic
design notation
diffusion of treatment
double-blind experiment
equivalent time-series design
experimental design
experimental group
experimental mortality
experimental realism
experimenter expectancy
external validity
factorial design
field experiment
hawthorne effect
history effects
interaction effect
internal validity
interrupted time-series design
laboratory experiment
latin square design
manipulation check
maturation effect
mundane realism
naturalistic generalization
one-shot case-study design
placebo effect
posttest
preexperimental designs
pretest
quasi-experimental designs
random assignment
reactivity
selection bias
solomon four-group design
static group comparison design
statistical regression effect
subjects
testing effect
treatment

REVIEW QUESTIONS

1. What are the seven elements or parts of an experiment?
2. What distinguishes preexperimental designs from the classical design?
3. Which design permits the testing of different sequences for several treatments?
4. A researcher says, "It was a three by two design with the independent variables being the level of fear (low, medium, high) and ease of escape (easy/difficult) and the dependent variable being anxiety." What does this mean? What is the design notation, assuming that random assignment with a posttest only was used?

5. How do the interrupted and the equivalent time series designs differ?
6. What is the logic of internal validity, and how does the use of a control group fit into that logic?
7. How does the Solomon four-group design show the testing effect?
8. What is a double-blind experiment, and why is it used?
9. Do field or laboratory experiments have higher internal validity? External validity? Explain.
10. What is the difference between experimental and mundane realism?

NOTES

1. Cook and Campbell (1979:9–36, 91–94) argued for a modification of a more rigid positivist approach to causality for experimental research. They suggested a "critical-realist" approach, which shares some features of the critical approach outlined in Chapter 4.
2. For discussions of the history of the experiment, see Danziger (1988), Gillespie (1988), Hornstein (1988), O'Donnell (1985), Scheibe (1988), and Webster and Sell (2007:6–9).
3. See Hornstein (1988:11).
4. For events after World War II, see Harris (1988) and Suls and Rosnow (1988). For a discussion of the increased use of deception, see Reynolds (1979:60).
5. See Field and Hole (2003) for a review of different comparisons.
6. Cook and Campbell (1979:7–9) and Spector (1981:15–16) discuss control in experiments.
7. The notation for research design is discussed in Cook and Campbell (1979:95–96), Dooley (1984:132–137), and Spector (1981:27–28).
8. For additional discussions of threats to internal validity, see Cook and Campbell (1979:51–68), Kercher (1992), Spector (1981:24–27), Smith and Glass (1987), and Suls and Rosnow (1988).
9. This example is borrowed from Mitchell and Jolley (1988:97).
10. Experimenter expectancy is discussed in Aronson and Carlsmith (1968:66–70), Dooley (1984:151–153), and Mitchell and Jolley (1988:327–329).
11. For discussions of external validity, see Aronson and Carlsmith (1968:22–25), Cook and Campbell (1979: 70–80), Lucas (2003), and Zelditch (2007).
12. For a discussion of external validity, see Lucas (2003), Mook (1983), Willer and Walker (2007b), and Vissersi et al. (2001).
13. 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 and how it shaped them.
14. See Piliavin and associates (1969).
15. See Hegtvedt (2007) for a recent review of ethical issues in experiments.
16. See Graham (1992).