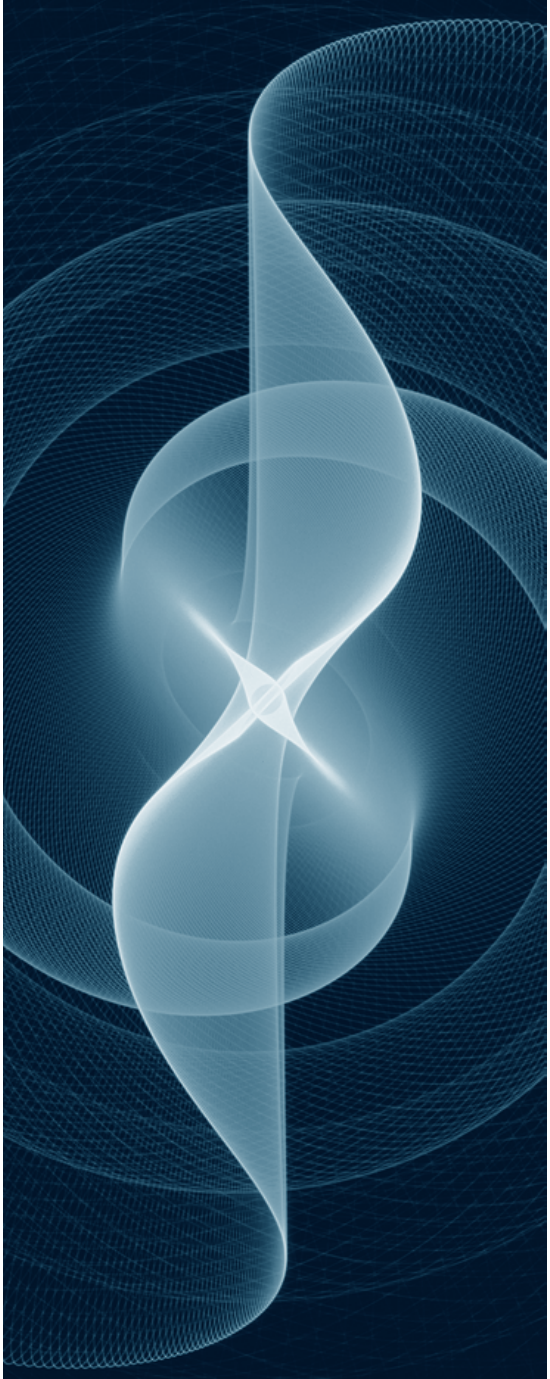
CHAPTER 3



RESEARCH TECHNIQUES: EXPERIMENTS

No one believes an hypothesis except its originator, but everyone believes an experiment except the experimenter. (W. I. B. BEVERIDGE)

Imagine you are a student in a class in environmental psychology and have received the following assignment: Go to the library and "defend" a table by preventing anyone else from sitting down for as long as you can. You must use only nonverbal and nonviolent means to accomplish this. To carry out this task, you might wait in the crowded library until a table is vacant, quickly sit down, and proceed to strew your books, clothing, and other belongings all over the table in hopes that this disarray might keep others away. After some time, say, fifteen minutes or so, someone finally does sit down at your table, ending your assignment. Have you performed an experiment?

Before answering this question, let us sketch out the major criteria for an experiment, which were briefly discussed in the preceding chapters. An experiment occurs when the environment is systematically manipulated so that the causal effect of this manipulation on some behavior can be observed. Aspects of the environment that are not of interest, and hence not manipulated, are held constant, so as not to influence the outcome of the experiment. We can then conclude that the behavior resulted from the manipulation. We must explain two special terms briefly introduced in Chapter 1—*independent* and *dependent* variables—to describe how the environment is manipulated and how behavior is observed.

WHAT IS AN EXPERIMENT?

Many students are surprised to discover that the actions described in our library table exercise do not constitute an experiment. All experiments require at least these two special features, the independent and dependent variables just mentioned. The **dependent variable** is the response measure of an experiment that is *dependent* on the subject. In this case, the time that elapsed until someone else sits down at the table is the dependent variable or response measure. The **independent variable** is a manipulation of the environment controlled by the experimenter: In this case, it is the strewing of articles on the table.

But an experiment must have at least two values, or **levels,** of the environment. These levels may differ in a quantitative sense (items strewn across only a portion of the table versus items strewn across the entire table), or the levels may reflect a qualitative difference (the person defending the table assumes a friendly, inviting expression as opposed to a stern, forbidding expression). The point is that at least two conditions must be compared with each other to determine if the independent variable (portion of table covered or facial expression) produces a change in a behavior or outcome. Sometimes, these two levels might simply be the presence or absence of a manipulation. The library example fails to meet this criterion, since it involves only one level of the independent variable.

How might we change the procedure to obtain an experiment? The simplest way would be to sit down again, this time without scattering anything. Then our independent

variable would have the necessary two levels: the table with items strewn about and the bare table with no items strewn about. Now we have something to compare with the first condition.

This experiment has three possible outcomes: (1) Strewing articles on the table results in a longer time before the table is invaded by another person; (2) the time until invasion is the same, whether or not articles are strewn about; and (3) scattering articles results in a shorter time until invasion. Without the second level of the independent variable (the table with no articles strewn about), these three outcomes cannot be formulated. Indeed, it is impossible to say anything about how effective articles strewn about are in defending library tables until two levels of the independent variable are tested.

When this library experiment is performed properly, the first possible outcome is obtained. A table can be better protected by a person plus assorted articles than by a person alone.

We can see, then, that experiments must have at least independent and dependent variables. The research techniques discussed in the preceding chapter do not allow or require manipulation of the environment; but before an **experiment** can be established, independent variables with at least two levels are necessary.

Advantages of Experiments

The main advantage of experiments over the techniques discussed in Chapter 2 is better control of extraneous variation. In the ideal experiment, no factors (variables) except the one being studied are permitted to influence the outcome; in the jargon of experimental psychology, we say that these other factors are *controlled*. If, as in the ideal experiment, all factors but one (that under investigation) are held constant, we can logically conclude that any differences in outcome must be caused by manipulation of that one independent variable. As the levels of the independent variable are changed, the resulting differences in the dependent variable can occur only because the independent variable has changed. In other words, changes in the independent variable cause the observed changes in the dependent variable. In the library example, we might want to manipulate the facial expression of the person "defending" the table. To control for extraneous variation, we would need to give careful consideration to other factors that might compromise our ability to make statements about causation. In this case, we might want to hire only one assistant to defend the table during the duration of the experiment or else establish objectively that our assistants are, for example, equally attractive. We might also decide to *control* for gender by either incorporating it as an additional independent variable or by using only female (or male) research assistants. Designing experiments so that there can be only one explanation of the results is at the heart of the experimental method. Whereas nonexperimental research techniques are limited to statements about description and correlation, experiments permit statements about causation—that is, independent variable A (facial expression) causes variable B (time elapsed until someone else sits down) to change. In this experiment, we would expect the time elapsed to be shorter when the assistant assumes a friendly and inviting expression than when the assistant's expression is stern and forbidding.

Thus, in principle, experiments lead to statements about causation. In practice, these statements are not always true. No experiment is 100 percent successful in eliminating or holding constant all other sources of variation but the one being studied.

However, experiments eliminate more extraneous variation than do other research techniques. Later in this chapter, we discuss specific ways in which experiments limit extraneous variation.

Another advantage of experiments is economy. Using the technique of naturalistic observation requires that the scientist wait patiently until the conditions of interest occur. If you lived in Trondheim, Norway—near the Arctic Circle—and wanted to study how heat affects aggression, relying on the sun to produce high temperatures would require great patience and lots of time. The experimenter controls the situation by creating the conditions of interest (various levels of heat in a laboratory setting), thus obtaining data quickly and efficiently.

Why Experiments Are Conducted

The same general reasons that apply to the conduct of any research also explain why psychologists perform experiments. In basic research, experiments are performed to test theories and to provide a database for explanations of behavior. These kinds of experiments are typically well planned, with the investigator having a clear idea of the anticipated outcome. So-called **critical experiments** try to pit against each other two theories that make different predictions. One outcome favors theory A; the other, theory B. Thus, in principle, the experiment will determine which theory to reject and which to keep. In practice, these critical experiments do not work out so well, because supporters of the rejected theory are ingenious in thinking up explanations to discredit the unfavorable interpretation of the experiment. One example of such an explanation is found in a study of how people forget. Two major explanations of forgetting are that (1) items decay or fade out over time, just the way an incandescent light bulb fades when the electricity is turned off (this explanation is called "trace decay") or that (2) items never fade, but because of this, they interfere with each other, causing confusion. A simple critical experiment would vary the time between introduction into memory of successive items, holding the number of items constant (Waugh & Norman, 1965). Memory should be worse with longer times, according to trace-decay theorists, because there is more time for items to fade out. But because the number of items remains the same regardless of the time at which they are introduced, interference theory predicts no differences in forgetting. When this experiment is performed, there is no difference in memory; this would seem to nullify the trace-decay explanation. The rejoinder by trace-decay theorists, however, is that the extra time given between items allows people to rehearse—that is, repeat the item to themselves—which prevents forgetting.

Less often, researchers perform an experiment in the absence of a compelling theory just to see what happens; we can call this a **what-if experiment.** Students often come up with what-if experiments, since these experiments require no knowledge of theory or the existing database and can be formulated on the basis of personal experience and observations. Some scientists frown on what-if experiments; the main objection to them is their inefficiency. If, as is often the case, nothing much happens in a what-if experiment—say, the independent variable has no effect—nothing is gained from the experiment. By contrast, if nothing much happens in a careful experiment for which a theory predicts something will happen, the finding of no difference can be useful. We must admit to having tried what-if experiments. Most of them did not work, but they were fun. Our advice is to check with your instructor before trying a what-if

experiment. He or she probably can give you an estimate of the odds of your coming up with anything or may even know the results of a similar experiment that has already been performed.

This brings us to the last major reason for doing experiments in basic research, which is to repeat or replicate a previous finding. A single experiment by itself is far less convincing than a series of related experiments. The simplest replication is the direct repetition of an existing experiment, with no change in procedures. Direct replications are especially useful when the original experiment was quite novel. Generally, however, a better way to replicate is to *extend* the previous procedure by adding something new while retaining something old. Thus, part of the replication is a literal repetition, but the novel part adds to scientific knowledge. This kind of repetition demonstrates the generality of a result by showing how it is (or is not) maintained over different independent variables. The concept of replication and its various forms are discussed at greater length in Chapter 11.

VARIABLES

Variables are the gears and cogs that make experiments run. Effective selection and manipulation of variables make the difference between a good experiment and a poor one. This section covers the three kinds of variables that must be carefully considered before starting an experiment: *independent, dependent,* and *control variables*. We conclude by discussing experiments that have more than one independent or dependent variable.

Independent Variables

In true experiments, independent variables are those *manipulated* by the experimenter. The brightness of a light, the loudness of a tone, the temperature of a room, the number of food pellets given to a rat—all are independent variables, since the experimenter determines their quality and quantity. Independent variables are selected because an experimenter believes they will cause changes in behavior. Increasing the intensity of a tone should increase the speed with which people respond to the tone. Increasing the number of pellets given to a rat for pressing a bar should increase the number of times the bar is pressed. When a change in the level (amount) of an independent variable causes a change in behavior, we say that the behavior is under control of the independent variable.

Failure of an independent variable to control behavior, often called **null results,** can have more than one interpretation. First, the experimenter may have guessed incorrectly that the independent variable was important: The null results may be correct. Most scientists will accept this interpretation only reluctantly, and so the following alternate explanations of null results are common. The experimenter may not have created a valid manipulation of the independent variable. Let us say you are conducting an experiment on second-grade children and your independent variable is the number of small candies (M&Ms, jelly beans) they get after each correct response. Some children get only one, whereas others get two. You find no difference in behavior. However, if your independent variable had involved a greater range—that is, from one piece of candy to ten pieces of candy—perhaps you would have obtained a

difference. Your manipulation might not have been sufficient to reveal an effect of the independent variable. Or perhaps, unknown to you, the children had a birthday party just before the experiment started and their little tummies were filled with ice cream and cake. In this case, maybe even ten pieces of candy would not show any effect. This is why, in studies of animal learning with food as a reward, the animals are deprived of food before the experiment starts.

We can see that experimenters must be careful to produce a strong manipulation of the independent variable. Failure to do so is a common cause of null results. Because there is no way to determine if the manipulation failed or the null results are correct, experimenters cannot reach any conclusions regarding the effect of the independent variable on the dependent variable. Other common causes of null results are related to dependent and control variables, to which we now turn.

Dependent Variables

The *dependent variable* is the response measure of an experiment that is *dependent* on the subject's response to our manipulation of the environment. In other words, the subject's behavior is observed and recorded by the experimenter and is dependent on the independent variable. Time elapsed before a subject sits down at a table defended by a research assistant, the speed of a worm crawling through a maze, the number of times a rat presses a bar—all are dependent variables, because they are dependent on the way in which the experimenter manipulates the environment. In the library example, we might predict that a subject would be more reluctant to sit down at a table that is defended by an assistant who displays a forbidding expression than if the assistant assumes a congenial expression. In this instance, the subject's behavior is *dependent* on the expression that we instruct the assistant to adopt. The time that elapses until the subject sits down at the table is the dependent variable of interest.

One criterion for a good dependent variable is **stability.** When an experiment is repeated exactly—same subject, same levels of independent variable, and so on—the dependent variable should yield the same score as it did previously. Instability can occur because of some deficit in the way we measure some dependent variable. Assume that we wish to measure the weight in grams of an object—say, a candle—before and after it is lit for 15 minutes. We use a scale that works by having a spring move a pointer. The spring contracts when it is cold and expands when it is hot. As long as our weight measurements are taken at constant temperatures, they will be reliable. But if temperature varies while objects are being weighed, the same object will yield different readings. Our dependent variable lacks stability.

Null results can often be caused by inadequacies in the dependent variable, even if it is stable. The most common cause is a restricted or limited range of the dependent variable, so that it gets "stuck" at the top or bottom of its scale. Imagine that you are teaching a rather uncoordinated friend how to bowl for the first time. Since you know from introductory psychology that reward improves performance, you offer to buy your friend a beer every time he or she gets a strike. Your friend gets all gutter balls, so you drink the beer yourself. Thus, you can no longer offer a reward; you therefore expect a decrement in performance. But since it is impossible to do any worse than all gutter balls, you cannot observe any decrement. Your friend is already at the bottom of the scale. This is called a **floor effect.** The opposite problem, getting 100 percent correct, is

called a **ceiling effect.** Ceiling and floor effects (see Chapter 10) prevent the influence of an independent variable from being accurately reflected in a dependent variable.

Control Variables

A **control variable** is a potential independent variable that is held constant during an experiment because it is controlled by the experimenter. For any one experiment, the list of relevant control variables is quite large, far larger than can ever be accomplished in practice. In even a relatively simple experiment—for example, requiring people to memorize three-letter syllables—many variables should be controlled. Time of day changes your efficiency; ideally, this should be controlled. Temperature could be important, because you might fall asleep if the testing room were too warm. Time since your last meal might also affect memory performance. Intelligence is also related. The list could be extended. In practice, an experimenter tries to control as many salient variables as possible, hoping that the effect of uncontrolled factors will be small relative to the effect of the independent variable. Although it is always important to exercise strict control over extraneous factors, it is even more critical when the independent variable produces a small effect on the dependent variable. Holding a variable constant is not the only way to remove extraneous variation. Statistical techniques (discussed later in the chapter) also control extraneous variables. However, holding a variable constant is the most direct experimental technique for controlling extraneous factors, so we limit our definition of control variables to only this technique. Null results often occur in an experiment because there is insufficient control of these other factors—that is, they have been left to vary systematically with the independent variable. Depending on the relationship between an extraneous variable and an independent variable, this uncontrolled variation can either obscure or inflate the effect of the independent variable on the dependent variable of interest. The problem of extraneous variation occurs more often in studies that are conducted outside of laboratories, where the ability to hold control variables constant is greatly decreased.

INDEPENDENT variable is MANIPULATED
DEPENDENT variable is OBSERVED
CONTROL variable is held CONSTANT

Name the Variables

Because understanding independent, dependent, and control variables is so important, we have included some examples for your use in checking your understanding. For each situation, name the three kinds of variables. The answers follow the examples. No peeking!

1. An automobile manufacturer wants to know how bright brake lights should be to minimize the time required for the driver of a following car to realize that the car in front is stopping. An experiment is conducted to answer this. Name the variables.

2. A pigeon is trained to peck a key if a green light is illuminated but not if a red light is on. Correct pecks are rewarded by access to grain. Name the variables.

3. A therapist tries to improve a patient's image of himself. Every time the patient says something positive about himself, the therapist rewards this by nodding, smiling, and being extra-attentive. Name the variables.

4. A social psychologist does an experiment to discover whether men or women give lower ratings of discomfort when six people are crowded into a telephone booth. Name the variables.

ANSWERS

1. Independent (manipulated) variable:	Intensity (brightness) of brake lights
Dependent (observed) variable:	Time from onset of brake lights until depression of brake pedal by following driver
Control (constant) variables:	Color of brake lights, shape of brake pedal, force needed to depress brake pedal, external illumination, etc.
2. Independent variable:	Color of light (red or green)
Dependent variable:	Number of key pecks
Control variables:	Hours of food deprivation, size of key, intensity of red and green lights, etc.
3. Independent variable:	Actually, this is not an experiment, because there is only one level of the independent variable. To make this an experiment, we need another level— say, rewarding positive statements about the patient's mother-in-law and ignoring negative ones. Then the independent variable would be: Kind of statement rewarded.
Dependent variable:	Number (or frequency) of statements
Control variables:	Office setting, therapist
4. Independent variable:	Gender of participant[1]
Dependent variable:	Rating of discomfort
Control variables:	Size of telephone booth, number of persons (six) crowded into booth, size of individuals, etc.

[1] Gender is a special type of independent variable called a subject variable, discussed later in this chapter.

More Than One Independent Variable

It is unusual to find an experiment reported in a psychological journal in which only one independent (manipulated) variable is used; the typical experiment manipulates from two to four independent variables simultaneously. This procedure has several advantages. First, it is often more efficient to conduct one experiment with, say, three independent

variables than to conduct three separate experiments. Second, experimental control is often better, since with a single experiment, some control variables—time of day, temperature, humidity, and so on—are more likely to be held constant than with three separate experiments. Third, and most important, is that results generalized—that is, shown to be valid in several situations—across several independent variables are more valuable than data that have yet to be generalized. Just as it is important to establish generality of results across different types of experimental subjects (see Chapter 12), experimenters also need to discover if some result is valid across levels of independent variables. Fourth, this allows us to study interactions, the relationships among independent variables. We illustrate these advantages with some examples.

Let us say we wish to find out which of two kinds of rewards facilitates the learning of geometry by high school students. The first reward is an outright cash payment for problems correctly solved; the second reward is early dismissal from class—that is, each correct solution entitles the student to leave class five minutes early. Assume that the results of this (hypothetical) experiment show early dismissal to be the better reward. Before we make early dismissal a universal rule in high school, we should first establish its generality by comparing the two kinds of reward in other classes, such as history or biology. Here, subject matter of the class would be a second independent variable. It would be better to put these two variables into a single experiment than to conduct two successive experiments. This would avoid problems of control, such as one class being tested the week of the big football game (when no reward would improve learning) and the other class being tested the week after the game is won (when students felt better about learning).

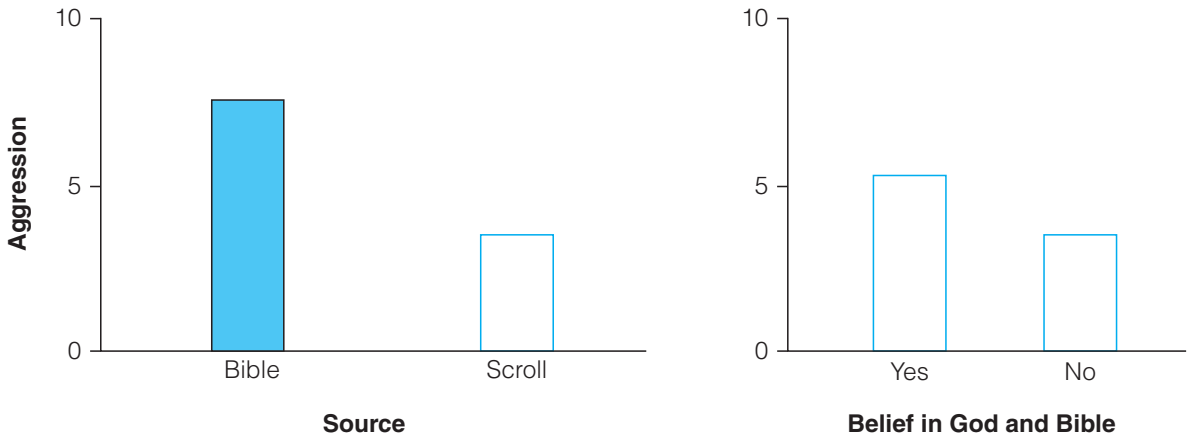
When the effects produced by one independent variable are different at each level of a second independent variable, we have an **interaction.** The search for interactions is a major reason for using more than one independent variable per experiment. This can best be demonstrated by example.

In a research report titled "When God Sanctions Killing," Bushman, Ridge, Da, Key, and Busath (2007) described a laboratory study of aggression. Participants read a violent passage that purportedly came from either the Bible or an ancient scroll. Following that, they performed an additional task that allowed them to present loud sounds to another subject in the experiment. They controlled the intensity of this sound, and higher intensities were interpreted as revealing greater aggression. The dependent variable was the number of times participants selected the highest noise levels in a set of 25 trials. Therefore, aggression scores could range from a low of 0 to a high of 25.

There were two independent variables. The first was the source of the violent passage: either the Bible or an ancient scroll. The second independent variable was whether or not the subject believed in God; this is a special type of independent variable, called a subject variable, which is discussed later in this chapter.

Results from this experiment are shown in Figure 3.1, with each independent variable plotted by itself. Reading a passage from the Bible produced greater aggression. Subjects who believed in God also acted more aggressively.

Figure 3.2 shows that this simple interpretation of the results, while correct, is incomplete. Here both independent variables are plotted on the same graph, making some relationships easier to see. If there was no mention of God because the passage came from an ancient scroll, subjects who believe in God and subjects who do not believe in God exhibited similar levels of aggression. But when God sanctioned violence because the passage came from the Bible, greater levels of aggression were exhibited by those subjects who believe in God.



▼ **FIGURE 3.1**

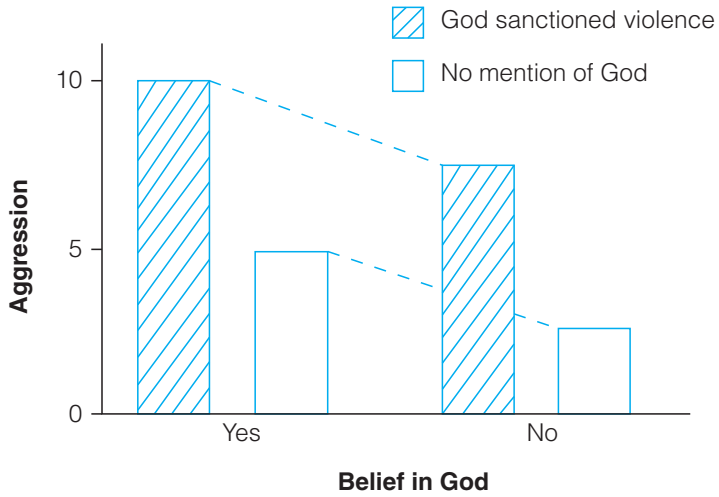
Effects of Two Independent Variables on Aggression. (Data from Bushman et al., 2007. Reprinted by permission of Blackwell Publishing.)

Remember, an interaction between two independent variables indicates that effects produced by one independent variable (belief in God) are not the same at each level of a second independent variable (source of the passage). When the passage contains no mention of God, belief in God has no effect upon aggression. But when the passage comes from the Bible, the increase in aggression is greater for subjects who believe in God than for subjects who do not believe in God. This is an interaction.

Figure 3.3 shows hypothetical data we invented to illustrate how these results might look if there were no interaction. The effect of one independent variable is the same at each level of the other independent variable. The dotted lines in Figure 3.3 are parallel, which is an easy way to detect the lack of an interaction. If similar lines were drawn in Figure 3.2., they would not be parallel because that figure shows an interaction of two independent variables.

Many experiments include two or more independent variables; this means that the results may contain an interaction. Because of the frequency with which you are likely to encounter interactions, we present another example of a two-variable experiment to help you practice interpreting the results of complex experiments.

In the experiment on social loafing (see Chapter 1) by Brickner, Harkins, and Ostrom (1986), the authors wanted to determine the effect of personal involvement in a task on the amount of social loafing shown on that task. Brickner and her associates noted that low-involvement tasks, such as clapping and generating uses for a knife, had been used in earlier research on social loafing. The authors reasoned that the effort devoted to a task should be related to the intrinsic importance or personal significance that the task has for the individual. High personal involvement in a task should reduce social loafing, because individuals should put forth a substantial amount of effort on such tasks, regardless of whether their individual performance is monitored. So, the researchers varied the subjects' involvement in the task and also varied the amount that individual effort could be assessed. If their reasoning was correct, there should be an interaction: Low involvement should lead to social loafing (reduced effort when the individual's effort cannot be assessed), but high involvement should lead to about the same amount of effort, whether or not individual effort could be identified.



▼ **FIGURE 3.3**

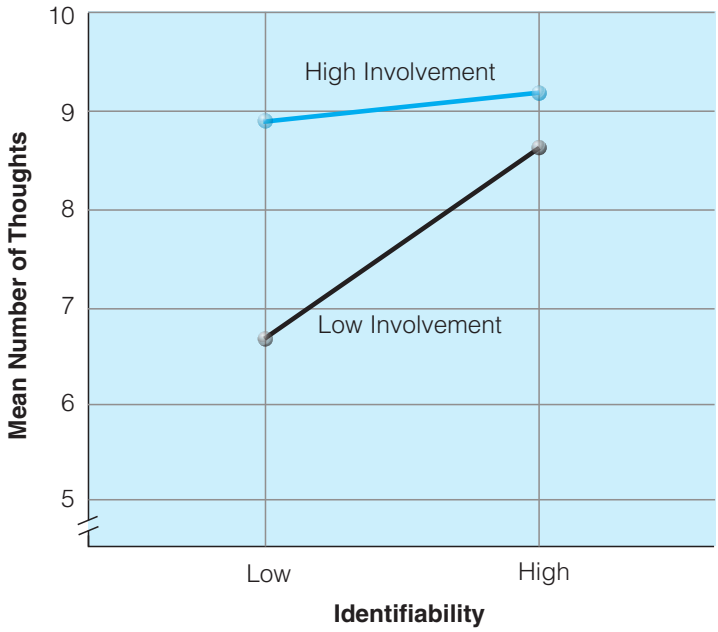
Hypothetical Data with No Interaction. Note Parallel Lines.

Brickner and associates had college students generate as many thoughts as they could in a 12-minute period about a proposal to implement senior comprehensive exams, which a student would have to pass in order to graduate. In the high-involvement condition, the students were led to believe that the proposal would be instituted at their college prior to their graduation.

Thus, the addition of comprehensive exams as one prerequisite to graduation should have high personal relevance. In the low personal-involvement condition, the students were led to believe that the exams would be instituted later, at another college. The possible identifiability of individual effort was also manipulated by instructions. Subjects wrote each of their thoughts about comprehensives on an individual slip of paper. In the low-identifiability condition, the subjects were told that their thoughts would be collected together with those of other subjects, because the committee evaluating the thoughts wanted to assess the range of opinions for the group as a whole. In the high-identifiability condition, the subjects were told that their opinions would be considered separately from those of others, because the committee in charge wanted to assess individual responses.

To summarize, the dependent variable was the number of thoughts generated in the four conditions: low identifiability and low involvement; low identifiability and high involvement; high identifiability and low involvement; and high identifiability and high involvement.

The results are shown in Figure 3.4, which plots the number of thoughts generated against identifiability for the two involvement conditions. Earlier social loafing research is replicated in the low-involvement condition: Fewer thoughts were generated when



▼ **FIGURE 3.4**

Results of the Experiment by Brickner, Harkins, and Ostrom (1986), Showing an Interaction. Social loafing (low numbers of thoughts generated with low as opposed to high identifiability) occurs with a low-involvement task but not with a high-involvement one.

the subjects believed that their individual performance was not being assessed. Now examine the results when there was high involvement: The number of thoughts was about the same, regardless of identifiability. Thus, the variables interact: The effects of identifiability depend on the level of task involvement. Put another way, social loafing, and therefore diffusion of responsibility, is less likely to occur when a person is confronted with a personally involving task than when the task does not have much intrinsic interest.

In summary, an interaction occurs when the levels of one independent variable are differentially affected by the levels of other independent variables. When interactions are present, it does not make sense to discuss the effects of each independent variable separately. Because the effects of one variable also depend on the levels of the other variables, we are forced to discuss interacting variables together.

More Than One Dependent Variable

The dependent (observed) variable is used as an index of behavior. It indicates how well or poorly the subject is performing. It permits the experimenter to score behavior. The experimenter must decide which aspects of behavior are relevant to the experiment at hand. Although some variables traditionally have been used, this does not mean that they are the only, or even the best, indexes of behavior. Take, for example, the behavior of a rat pressing a bar or a pigeon pecking a key, responses that are used in studies of animal learning. The most common dependent variable is the number of presses or pecks observed. But the force with which a key is pecked can also lead to interesting findings (see Neuringer 2002, p. 680; Notterman & Mintz, 1965), as can the latency (the time taken to respond). Researchers can usually come up with several dependent variables that may be appropriate. Let us say we wish to study the legibility of the typeface that you are now reading. We cannot observe "legibility," of course. What dependent variables might we observe? Here are some that have been used in the past: retention of meaningful information after reading text, time needed to read a fixed number of words, number of errors in recognizing single letters, speed in transcribing or retyping text, heart rate during reading, and muscular tension during reading—and this list is far from exhaustive.

Reasons of economy argue for obtaining as many dependent measures at the same time as is feasible. Despite this, the typical experiment uses only one, or at the most, two dependent variables simultaneously. This is unfortunate: Just as the generality of an experiment is expanded by having more than one independent variable, it is also expanded with several dependent variables. The reason why more dependent variables are not used is probably because it is statistically difficult to analyze several dependent variables at once. Although modern computer techniques make the calculations quite feasible, many experimental psychologists have not been well trained in these multivariate statistical procedures and thus hesitate to use them. Separate analyses could be conducted for each dependent variable by itself, but this loses information in much the same way that a separate analysis of independent variables ignores interactions. Multivariate analysis is complex; nevertheless, you should be aware that it is often advantageous to use more than one dependent variable in an experiment.

EXPERIMENTAL DESIGNS

The purpose of experimental design is to minimize extraneous or uncontrolled variation, thereby increasing the likelihood that an experiment will produce valid, consistent results. Entire books have been written about experimental design. Here, we cover a sample of some common techniques used to improve the design of experiments.

One of the first design decisions an experimenter must make is how to assign subjects to the various levels of independent variables. The two main possibilities are to assign only some subjects to each level or to assign each subject to every level. The first possibility is called a **between-subjects design** and the second, a **within-subjects design.** The difference can be shown with a simple example. Thirty students in introductory psychology have signed up for an experiment that you are conducting to test ability to remember nonsense words. Your independent variable is the number of times you will say each item: one time or five times. You expect that an item presented five times will be learned better than an item presented only once. The between-subjects design calls for you to divide your subjects by halves—that is, into two groups of 15 students each—with one group receiving five repetitions and the other, one repetition. (How to select which subjects to put in each group is discussed shortly.) The within-subjects design has all 30 subjects learning with both levels of the independent variable—that is, each is tested with one repetition and again with five repetitions. (How to determine the order in which each subject gets these two treatments is also discussed later.) Which design should you use?

Between-Subjects Designs

The between-subjects (two groups) design is conservative. There is no chance that one treatment will continue to contaminate the other, because each person receives only one treatment (one repetition or five repetitions, but not both). One drawback, however, is that the between-subjects design must deal with differences among people, and this decreases its efficiency—that is, its ability to detect real differences between one and five repetitions of the memory items.

In any between-subjects design, the experimenter must try to minimize differences among the subjects in the two or more treatment groups. Clearly, if we took the five best memorizers and deliberately placed them in the one-repetition group, and put the five worst in the five-repetition group, we might wind up with no difference in results—even, perhaps, with the one-repetition group doing better. To prevent this outcome, the experimenter must ensure that both groups are equivalent at the start of the experiment.

Equivalent Groups One way to ensure equivalence would be to administer a memory test to all 30 subjects before the regular experiment started, to obtain a **baseline** measure of the subjects' ability to memorize nonsense words. Subjects' baseline scores could then be used to form pairs of subjects that had equal or very similar scores. One member of each pair would be randomly assigned to one group and the other member to the second group. This technique is called **matching**. One difficulty with matching is that an experimenter cannot match subjects on every possible characteristic. Thus, there is always the possibility that the groups, even though matched on

some characteristic(s), differ on some other characteristic that may be relevant (matching is discussed in greater detail later in this chapter).

A more common technique used to ensure that equivalent groups are formed is **randomization.** Randomization means that each person participating in an experiment has an equal chance of being assigned to any particular group. In our repetition experiment, one way to form two groups by randomization would be to draw names out of a hat. Or we could ask each person to step forward and then throw a die. Even throws would be assigned to one group and odd throws to the other. If we did not have any dice, a table of random numbers could be used to generate even and odd digits. This method of assigning subjects to experimental conditions has no bias, since it ignores all characteristics of the subjects; we expect that the groups so created would be equivalent on any and all relevant dimensions. However, randomization does not guarantee that groups will always be equal. By chance, a greater number of better memorizers might be assigned to one of the groups. The odds of this occurring can be calculated by the methods of probability theory as applied to statistics (see Appendix B). This is one reason why experimental designs and statistics are often treated as the same topic. However, design is concerned with the logic of arranging experiments, whereas statistics deals with calculating odds, probabilities, and other mathematical quantities.

If we are sure that all relevant dimensions have been dealt with, matching is preferable to randomization. But because we seldom are sure, randomization is used more often.

Within-Subjects Designs

Many experimental psychologists would prefer the within-subjects (one group) design in which all 30 subjects were tested with one repetition and again with five repetitions (or vice versa). It is more efficient, since each subject is compared with himself or herself. Any differences resulting from one versus five repetitions cannot be the result of differences between the people in the two groups, as might be the case for the between-subjects design.

General Practice Effects There is a risk, however, in the more-efficient within-subjects design. Imagine that all 30 subjects first learn a large number of items with five repetitions and then learn with one repetition. By the time subjects begin the one-repetition treatment, they might have become more proficient in learning nonsense words, or they might be experiencing some boredom or fatigue with the task. Both these possibilities are termed **general practice effects.** These effects are usually assumed to be the same for all treatment conditions so that it does not matter whether subjects learned with one repetition followed by five repetitions or five followed by one repetition. Because general practice effects are the same for all treatment conditions, they can be controlled largely through **counterbalancing.** With counterbalancing, the experimenter faces the difficulty of determining the order in which treatments should be given to subjects. Again, one solution is to use randomization by drawing the treatment titles out of a hat, using a random-number table, or using a computer to order conditions randomly. The logic behind this was discussed earlier. However, although counterbalancing treatments through randomization produces equivalent orders in the long run, it is less likely to be suitable when there are only a small number of treatments. In most

experiments, the number of subjects exceeds the number of treatments, so randomization is a good technique for assigning subjects to treatments.

Complete counterbalancing makes sure that all possible treatment orders are used. In the repetition experiment, this is easy because there are only two orders: one and five repetitions, five and one repetitions. Half the subjects would receive one repetition followed by five repetitions, and the other half would get the opposite order. As the number of treatments increases, the number of orders becomes large indeed. Three treatments have 6 different orders; four treatments have 24 different orders; five treatments have 120 different orders; and so on. As the levels of an independent variable increase, complete counterbalancing soon becomes impractical.

Counterbalancing does not eliminate the effects of order. It does allow experimenters to evaluate possible order effects. If such effects are present, and especially if they form interactions with other, more important independent variables, steps need to be taken to correct the design. The experimenter might decide to repeat the experiment, using a between-subjects design to avoid order effects. Alternatively, the original experiment could be reanalyzed as a between-subjects one, by examining behavior in just the initial condition experienced by each subject.

Differential Carryover Effects Differential carryover effects pose a more serious problem than do general practice effects. In the case of **differential carryover effects,** the effect of the early part of the experiment on the later part of the experiment varies depending on which treatment comes first. Imagine that all 30 subjects first learn items with five repetitions and then learn with one repetition. As a result of their earlier experience with five repetitions, they might decide to repeat to themselves four more times the item that was only presented once. This would destroy any differences between the two levels of the independent variable. This is an example of a differential carryover effect given that the effect of the first treatment on the second treatment differs depending on which treatment came first. This was not the case with general practice effects in which subjects approached the second treatment in the same way (i.e., with greater skill, boredom, or fatigue), regardless of the treatment they received in the first phase of the experiment. Differential carryover effects can be diminished somewhat through counterbalancing, but counterbalancing cannot eliminate these effects entirely. If there is reason to expect differential carryover effects, we can do one of two things in addition to counterbalancing: use the between-subjects design or build in a sufficient time delay between the two treatments. Because the between-subjects design is less efficient, it will require that many more subjects be tested; but this is preferable to conducting a seriously flawed experiment. If we decide to insert a time delay between the two treatments, we must identify a duration of time that is sufficient to eliminate the possibility of differential carryover effects.

Small-*n* Designs

Before turning to a discussion of mixed designs, we would like to mention a variant on the traditional within-subjects design—the **small-*n* design.** Small-*n* designs present the levels of the independent variable or treatments to a small number of subjects or a single subject. Because few subjects are tested, a substantial number of observations are recorded for each subject, resulting in a very economical and highly controlled experiment. Small-*n* experiments are common in psychophysical, clinical, and operant-conditioning

research. Just as with the within-subjects design, the experimenter must be careful to counterbalance treatments and anticipate any problems associated with administering multiple treatments to individual subjects. Small-*n* designs are discussed at length in Chapter 9 of this text and in Chapter 9 of Elmes, Kantowitz, and Roediger (2003).

Mixed Designs

Experiments need not be exclusively of within-subjects or between-subjects design. It is often convenient and prudent to have some independent variables treated as between-subjects and others as within-subjects in the same experiment (assuming the experiment has more than one independent variable, of course). If one variable—for example, the administration of a drug—seems likely to affect others, it can be made a between-subjects variable, while the rest of the variables are varied within subjects. When trials or repeated practice on a task are of interest, it is of necessity a within-subjects variable. Frequently, a mixed design is used, in which some variable is imposed between subjects to see its effect across a second, within-subjects variable. This type of compromise design (**mixed design**) is not as efficient or economical as a pure within-subjects design, but it is often safer.

Control Conditions

Independent variables must be varied (or manipulated) by the experimenter. This implies that each and every independent variable must vary either in amount (quantitative variation) or in kind (qualitative variation) within the experiment. For example, if the amount of reward given to a rat is an independent variable, the amounts chosen by the experimenter might be one and four pellets of food. Alternatively, we could offer different kinds of rewards, such as food and water. The technical term for a single treatment or condition of an independent variable is level. We would state that the levels of the independent variable are one and four food pellets in the first example and food and water in the second example.

Many experiments contain, in addition to independent variables, some **control group** (between-subjects design) or **control condition** (within-subjects design). In its simplest form, the control group does not receive the levels of interest of the independent variable. In the reward example just described, a control group of rats would receive no reward. Or say an experimenter is interested in the effect of noise on studying. Using a between-subjects design, the experimenter would expose one group of subjects to loud noise for half an hour while they were studying; this is the level of interest of the independent variable. A control group would study the same material for half an hour in a quiet setting (a very low level of noise). Then both groups would be tested on the material. Any obtained difference on the test between the two groups would be attributed to the effect of noise.

The important characteristic of a control condition is the fact that it provides a baseline against which some variable of interest can be compared. Sometimes the best baseline is no treatment, but often the best baseline requires some activity. A frequent example occurs in memory research, where a group of subjects is required to learn two different lists of words. The experimenter is interested in how learning one list interferes with

Experimental Group	Learn List A	Learn List B	Test List A
Control Group	Learn List A	Do Arithmetic	Test List A

▼ **FIGURE 3.5**

Examples of Experimental and Control Groups for List Learning.

learning the other. The experimental group (receiving the level of interest of the independent variable) first learns list A, then learns list B, and then is tested again on list A. The experimenter would like to show that learning list B interferes with retaining list A. But before any conclusion of this sort can be reached, a comparison control condition is required. Merely comparing the final test of list A with the first test is insufficient, because subjects might do worse on the last list A test simply because they are tired, or they might do better because they have had extra practice. A control condition with no treatment would have a control group learn list A, then sit around for the time it took the experimental group to learn list B, and then be tested again on list A. But this would be a poor control condition, because subjects might practice or rehearse list A while they were sitting around. This would improve their final performance on the last list A test and incorrectly make it appear that in the experimental group, list B interfered more than it really did with list A. A proper baseline condition would occupy the control group during the time the experimental group was learning list B; perhaps the experimenter would have them do arithmetic or some other "busy work" that would prevent rehearsal (Figure 3.5).

Sometimes the control condition is contained implicitly within the experiment. Recall the memory experiment discussed earlier, in which the independent variable was the number of repetitions of an item: one or five. No experimenter would bother to include a control group or condition with zero repetitions, since no learning could occur under this odd circumstance. The control condition is implicit, in that five repetitions can be compared with one, and vice versa. Since the experimenter might well be as interested in the effects of a single repetition as in five repetitions, we probably would not explicitly call the one-repetition level a control condition. But it does provide a baseline for comparison—and so, for that matter, does the five-repetition condition, since the one-repetition results can be compared with it.

Many types of experiments require more than one baseline. In physiological and drug research, for example, a control for surgical or injection trauma is needed. So, a subject might receive a sham operation or the injection of an inert substance (a placebo) in the control condition; those would also be compared with other controls that received no operation or no injection.

Pitfalls

Unfortunately, it is quite easy to formulate an inadequate experimental design; most experimental psychologists have hidden away mistakes of this kind in a dusty file cabinet. In this section, we discuss only a small sample of errors in design, those that are so common you should be aware of them.

Demand Characteristics Laboratory experiments attempt to capture behavior as it really is influenced by the independent variable. Sometimes the laboratory setting itself or the knowledge that an experiment is under way may alter patterns of behavior. Many times, research participants spontaneously form hypotheses or assumptions about the experimenter's purpose in conducting the experiment and then behave or respond in a way that will satisfy this "purpose." Try this simple demonstration to convince yourself that such effects occur. Tell five of your friends that you are conducting an experiment for your psychology class and would like their cooperation as subjects. If they agree, ask them to hold three ice cubes in their bare hands. Note how many hold the ice cubes until they melt. Now ask five other friends to hold the ice cubes, without mentioning anything about an experiment. Instead of holding the ice cubes until they melt, they will consider your request somewhat strange and soon so inform you. There is something unusual about the ready compliance of those friends who knew they were participating in an experiment: More of them were willing to hold the ice cubes for a longer period. Psychologists call the cues available to subjects that allow them to determine the purpose of the experiment, or what is expected by the experimenter, **demand characteristics.** To the extent that the behavior of research participants is controlled by demand characteristics instead of by independent variables, experiments are invalid and cannot be generalized beyond the test situation.

A well-known example of a demand characteristic is the **Hawthorne effect,** named after the Western Electric Company plant where it was first observed. The company was interested in improving worker morale and productivity and conducted several experiments (such as improving lighting) to better the workers' environment. No matter what experimental manipulation was tried, worker productivity improved. The workers knew they were in a "special" group, and therefore tried to do their best at all times. (See Bramel & Friend, 1981, and Parsons, 1974, for alternate interpretations of these results.) The demand characteristics were more important in determining the workers' productivity than were the experimental manipulations. Although the term *Hawthorne effect* is widely used to describe field experiments where productivity increased due to participation in the study, there have been several detailed reviews of the original Hawthorne experiment that suggest the original conclusion was based upon weak evidence (Brannigan & Zwerman, 2001; Wickström & Bendix, 2000). Nevertheless, the term remains in wide use.

Demand characteristics, and the Hawthorne effect, must be carefully evaluated. A recent study (Fostervold, Buckmann, & Lie, 2001) contained special control conditions for evaluating the effects of visual display unit (VDU) filters on computer screens. In the first part of the study one group of participants had filters (filter group) and another control group did not. Comparing the two groups' results showed various benefits for the filter group. However, the researchers also included a second phase where the control group was given a filter while the filter group continued with the same filter. Only minor changes were observed for the initial control group. Furthermore, initial benefits for the filter group declined during the second phase. Thus, results in the first phase were due to demand characteristics and not to benefits associated with VDU filters. Had the experimenters conducted only the first phase of their study, a false benefit of filters, actually due to demand characteristics, might have been claimed incorrectly.

Experimenter Effects A pitfall closely related to demand characteristics is the **experimenter effect,** which influences the outcome accidentally by providing participants with slight cues as to the experimenter's expectations. For example, an experimenter

might not be aware that he or she nods approvingly when a correct response is given and frowns after errors. The gender, race, and ethnicity of the experimenter are also potential experimenter effects. Experimenter characteristics are more likely to bias the results of an experiment in research that focuses on issues related to these characteristics—for example, the race of an experimenter who is conducting an experiment concerning the effect of skin color on work performance ratings.

These effects are not limited to experiments with humans. The experimenter effect can also occur in seemingly objective experiments with animal subjects. Rosenthal and Fode (1963) told student experimenters that the rats they were to test in a maze were from special strains: either maze-bright or maze-dull. Actually, the rats came from the same population. Nevertheless, the rats that were labeled maze-bright had fewer errors than those labeled maze-dull, and this difference was statistically reliable. The student experimenters were observed while they tested the rats: They did not cheat or do anything overt to bias the results. It seems reasonable that the lucky students who got supposedly bright rats were more motivated to perform the experiment than those unfortunates who had to teach stupid rats to go through the maze. Somehow, this affected the results of the experiment—perhaps because experimenters handled the two groups of rats differently.

The best way to eliminate this kind of experimenter effect is to hide the experimental condition from the experimenter on the premise that experimenters cannot communicate what they do not know. This procedure is termed a **double-blind experiment** because neither the experimenter nor the research participant knows which subjects are in which treatment conditions. Such a procedure was, for instance, used in a study of behavioral effects of air pollution. Subjects breathed either pure air or air taken from a busy roadway. The air was contained in tanks; the experimenter did not know which tank held pure air and which tank held polluted air. The subjects' poorer performance in polluted air cannot, then, be attributed to the experimenter inadvertently disclosing the air quality to subjects or treating them differently.

Experimenter effects are not always this subtle. One of the authors was once involved in an experiment concerning the human eye-blink response. Several experimenters helped conduct the same experiment, and it was soon noticed that one of them obtained results that were quite different from those of the rest of us. His subjects started out experimental sessions with massive flurries of frenzied blinking. The cause of this odd behavior was easily discovered. To record eye blinks, the experimenter must attach a tiny metal rod to the subject's eyelid with special tape—ordinarily a painless procedure. However, the experimenter in question had a very heavy thumb and was unable to attach the rod without irritating the eye, causing the strange flurries of blinking.

When an experimenter suspects that some aspect of his or her appearance or manner (e.g., gender, race, ethnicity) may alter the pattern of subjects' behavior, then a possible solution is to incorporate this as an additional independent variable or control variable in the experimental design. If an African-American experimenter is conducting research on skin color and work performance ratings, he or she could ask a white colleague or research assistant to test half the subjects and then compare the effects of skin color in the two experimenter race conditions.

Automation of Experiments Experimenter effects can be eliminated or greatly reduced by having computers or other equipment conduct the experiment so that the subject is untouched by human hands. In many laboratories, a subject enters a testing

booth and sees a message on a screen that tells her or him to push a button to begin. Pushing the button causes instructions for the experiment to appear on the screen. The entire experiment is then conducted by a computer. The experimenter appears at the end of the data collection to debrief the participant, giving the aims of the study and explaining how the subject has helped advance science. Until then, the experimenter simply monitors the equipment and the subject to ensure that the subject is following instructions and that nothing untoward happens. Such automation obviously reduces the dangers of experimenter bias.

Quasi-Experiments

For one reason or another, many variables cannot be manipulated directly. One deterrent to manipulation of variables in experiments is the ethical considerations all scientists must have (see Chapter 4). It is ethical to survey or otherwise observe the use of drugs by college students as long as permission is obtained. By no stretch of the imagination, however, would it be ethical to create a group of drug abusers and compare their activities with a nonabusing group that we also created. A second barrier to manipulation is Mother Nature. Some variables, such as the sex of our subjects, cannot be varied by the experimenter (except in very rare and controversial circumstances); other variables, such as natural disasters (tornadoes, hurricanes) or unnatural disasters (wars, airplane crashes), are both physically and morally difficult to implement. Can we do experiments that concern these phenomena? After all, such variables and others like them are fascinating and may play an important part in human experience.

The answer to the question (assuming you are an ethical scientist) is this: You can and you cannot. We are not being silly here; rather, we are emphasizing the fact that you cannot do real experiments on phenomena such as the ones just listed. You can, however, conduct **quasi-experiments.** The technique here is similar to the ex post facto examination in correlational research, except that two or more levels of the variable of interest are examined rather than correlated. We wait for Mother Nature to do her work, and then we compare the effects of that "independent variable" with the effects that occur when that variable is not present or differs in some way. If we compare the reading ability of men with that of women, or that of speed readers with that of average adults, we have conducted a quasi-experiment.

The advantages of quasi-experiments are obvious: They use naturally occurring independent variables, most of which have a high degree of intrinsic interest and important practical implications. In a quasi-experiment, we take advantage of observational and correlational procedures and combine them with the power of experimentation. The typical quasi-experiment has a **subject variable** as an independent variable. If we want to find out about almost any inherent subject variable (age, sex, race, ethnic group), socially caused subject attribute (social class, region of residence), or disease- and illness-related subject attribute (limb loss, mental illness, brain damage, effects of disasters), we are going to have to select rather than vary our independent variables, unless it is possible to do the experiment directly on infrahuman organisms. Although quasi-experiments are interesting and can contribute very important research, we should caution you here that the advantages of quasi-experiments are gained at the expense of control. When the researcher has to take what is given, what is given may include several important confounding variables.

Because much research in psychology is concerned with subject variables and because quasi-experiments using subject variables are likely to be confounded, we now examine the problems and possible solutions.

An experimenter cannot manipulate a subject variable while holding other factors constant; she or he can only select subjects who already have the characteristic in some varying degree and then compare them based on the behavior of interest. If the subjects in the different groups (say, high, medium, and low IQ) differ on the behavior, we cannot conclude that the subject-variable difference has produced or is responsible for the difference in behavior. The reason is that other factors may be covariant and confounded with the subject variable. If high-IQ subjects perform some task better than low-IQ subjects, we cannot say that IQ produced or caused the difference, because the different groups of subjects are likely to vary on other relevant dimensions, such as motivation, education, and so forth. When subject variables are investigated, we cannot safely attribute differences in behavior to this variable, as we can with true experimental variables. Such designs, then, essentially produce correlations between variables. We can say that the variables are related, but we cannot say that one variable produces or causes the effect in the other variable.

This is a very important point; let us consider an example. Suppose an investigator is interested in the intellectual functioning (or lack thereof) of people suffering from schizophrenia. People diagnosed as belonging to this group are given numerous tests that are meant to measure various mental abilities. The researcher also gives these tests to another group of people, so-called normals. He or she discovers that schizophrenics do especially poorly relative to normals in tests involving semantic aspects of language, such as those that involve understanding the meanings of words or comprehending prose passages. The investigator concludes that the schizophrenics perform these tests more poorly *because* they are schizophrenics and that their inability to use language well in communication is a likely contributing cause of schizophrenia.

Studies such as this are common in some areas of psychology. Despite the fact that conclusions similar to this are often drawn from such studies, they are completely unwarranted. Both conclusions are based on correlations, and other factors could well be the critical ones. Schizophrenics may do more poorly than normals for any number of reasons. They may not be as intelligent, as motivated, as educated, or as wise at taking tests. It may simply be that they have been institutionalized for a long time, with a resulting poverty of social and intellectual intercourse. So we cannot conclude that the reason that the two groups differ on verbal tests is schizophrenia or its absence in the two groups. Even if we could conclude this, it would certainly not imply the other conclusion, that language problems are involved in causing schizophrenia. Again, all we would have is a correlation between these two variables, with no idea of whether or how the two are causally related.

Use of subject variables is very common in all psychological research, but it is absolutely crucial in such areas as clinical and developmental psychology. Therefore, the problems with making inferences from such research should be carefully considered. A primary variable in developmental psychology is age, a subject variable; this means that much research in this field is correlational in nature. In general, the problem of individual differences among subjects in psychology is one that is often ignored, though there are often appeals to consider this problem as crucial (see Underwood, 1975). We devote a chapter later in the book to individual differences (Chapter 12). Let us consider here one way of attempting more sound inferences from experiments employing subject variables.

Matching Again The basic problem in the investigation of subject variables and in other ex post facto research is the fact that whatever differences are observed in behavior may be caused by their confounded variables. One way to try to avoid this problem is by matching subjects on the other relevant variables. In the comparison of schizophrenic and normal subjects, we noted that the two groups were also likely to differ on other characteristics, such as IQ, education, motivation, institutionalization, and perhaps even age. Rather than simply comparing the schizophrenic subjects with normal subjects, we might try to compare them with another group more closely matched on these other dimensions, so that, we hope, the main difference between the groups would be the presence or absence of schizophrenia. For example, we might use a group of patients who, on the average, are similar to the schizophrenics in terms of age, IQ, length of time institutionalized, gender, and some measure of motivation. When the two groups have been matched on all these characteristics, then we can more confidently attribute any difference in performance between them to the factor of interest, namely, schizophrenia. By matching, investigators attempt to introduce the crucial characteristic of experimentation—being able to hold constant extraneous factors to avoid confoundings—into what is essentially a correlational observation. The desire is to allow one to infer that the variable of interest (schizophrenia) produces the observed effect.

Several rather severe problems are associated with matching. For one thing, it often requires a great deal of effort, because some of the relevant variables may be quite difficult to measure. Even when one goes to the trouble of taking the needed additional measures, it may still be impossible to match the groups, especially if few subjects are involved before matching is attempted. Even when matching is successful, it often greatly reduces the size of the sample on which the observations are made. We then have less confidence in our observations, because they may not be stable and repeatable.

Matching is often difficult because crucial differences among subjects may have subtle effects. In addition, the effects of one difference may interact with another. Thus, *subtle interactions* among matched variables may confound the results. To illustrate these difficulties, let us consider some of the work done by Lester and Brazelton (1982) on neonatal behavior.

Brazelton's primary interest is in cultural differences in neonatal behavior, as measured by the Brazelton Neonatal Behavioral Assessment Scale. The general strategy is to compare neonates from various cultures and ethnic groups with neonates from the United States. In these quasi-experiments, culture or ethnic group, which is a subject variable, is the quasi-independent variable. Attempts are usually made to match the babies from different cultures along various dimensions, such as birth weight, birth length, and obstetrical risk (including whether the mother received medication during birth, whether the baby was premature, and so on). Lester and Brazelton show that there is a synergistic relationship among these factors. **Synergism** in a medical context means that the combined effects of two or more variables are not additive: The combined effect is greater than the sum of the individual components. This means that the variables interact.

The way in which neonatal characteristics and obstetrical risk interact is as follows. Studies have shown that the behavior (as measured by the Brazelton scale) of slightly underweight infants is more strongly influenced (negatively) by small amounts of medication taken by the mother than is the behavior of neonates who are closer to the average in weight. Even though the neonates are carefully selected, subtle and interactive effects of the matched variables can influence the results. This is an especially difficult problem in Brazelton's work, because much of his research has examined

neonates from impoverished cultures, where birth weight is low and obstetrical risk is very high. Generally, you should remember that matched variables are rarely under direct control, which means that the possibility of confounding is always present.

Another problem with matching involves the introduction of the dreaded **regression artifact.** This is discussed in Chapter 12, but we explain it briefly here. Under certain conditions in many types of measurements, a statistical phenomenon occurs known as **regression to the mean.** The mean of a group of scores is what most people think of as the average: the total of all observations divided by the number of observations. For example, mean height in a sample of 60 people is the sum of all their heights divided by 60. Typically, if people who received extreme scores (i.e., very high or very low) on some characteristic are retested, their second scores will be closer to the mean of the entire group than were their original scores. Consider an example. We give 200 people a standard test of mathematical reasoning for which there are two equivalent forms, or two versions of the test that we know to be equivalent. The average (mean) score on the test is 60 of 100 possible points. We take the 15 people who score highest and the 15 who score lowest. The mean of these groups is, say 95 and 30, respectively. Then we test them again on the other version of the test. Now we might find that the means of the two groups are 87 and 35. On the second test, the scores of these two extreme groups regress toward the mean; the high-scoring group scores more poorly, and the low-scoring group does somewhat better. Basically, this happens for the high-scoring group because some people whose "true scores" are somewhat lower than actually tested lucked out and scored higher than they should have on the test. When retested, people with extremely high scores tend to score lower, near their true score. The situation is reversed for the low-scoring group. That is, some of them scored below their "true scores" on the first test; retesting leads to their scoring higher or nearer the true score.

This regression toward the mean is always observed under conditions when there is a less-than-perfect correlation between the two measures. The more extreme the selection of scores, the greater the regression toward the mean. It also occurs in all types of measurement situations. If abnormally tall or short parents have a child, it will likely be closer to the population mean than the height of the parents. As with most statistical phenomena, regression to the mean is true of groups of observations and is probabilistic (i.e., it may not occur every time). For example, a few individual subjects may move away from the mean in the second test of mathematical reasoning, but the group tendency will be toward the mean.

How does regression toward the mean affect quasi-experiments, in which subjects have been matched on some variable? Again, consider an example. This one, like much ex post facto research done on applied societal problems, has important implications. Let us assume that we have an educational program that we believe will be especially advantageous for increasing the reading scores of African-American children. This is especially important because African-American children's scores are typically lower than those of whites, presumably because of different cultural environments. We take two groups of children, one African-American and one white, and match them on several criteria, including age, sex, and, most important, initial reading performance. We give both groups of children the reading improvement program and then test their reading scores after the program. We find, much to our surprise, that the African-American children actually perform worse after the reading program than before it, and the white children improve. We conclude, of course, that the program

helped white children but actually hurt African-American children, despite the fact that it was especially designed for the latter.

This conclusion, even though it may seem reasonable to you, is almost surely erroneous in this case, because of regression artifacts. Consider what happened when the African-American and white children were matched on initial reading scores. Since the populations differed initially, with African-Americans scoring lower than whites, in order to match two samples it was necessary to select the African-American students having higher scores than the mean for their group and the white students having lower scores than their group mean. Having picked these extreme groups, we would predict (because of regression to the mean) that when retested, the African-American children would have poorer scores and the white children would have better ones, on the average, even if the reading improvement program had no effect at all! The exceptionally high-scoring African-American children would tend to regress toward the mean of their group, and the low-scoring whites would regress toward the mean for their group. The same thing would have happened even if there had been no program and the children had been simply retested.

The same outcome would likely have been obtained if children had been matched on IQs instead of reading scores, since the two are probably positively correlated. So simply finding another matching variable may not be a solution. One solution would be to match very large samples of African-American and white children and then split each group, giving the reading program to one subgroup but not the other. All would be retested at the end of the one subgroup's participation in the program. (Assignment of subjects to the subgroups of African-American and white children should, of course, be random.) Regression to the mean would be expected in both subgroups, but the effect of the reading program could be evaluated against the group that had no program. Perhaps African-American children with the reading program would show much less drop (regression to the mean) than those without, indicating that the program really did have a positive effect.

Because quasi-experimental research with subject variables is conducted quite often to evaluate educational programs, its practitioners need to be aware of the many thorny problems associated with its use. One may not be able to say much with regard to the results or draw important conclusions because of confoundings. Matching helps alleviate this problem in some cases where its use is possible, but then one introduces the possibility of regression artifacts. And many researchers seem unaware of this problem. One famous blooper in such evaluational research, very similar to the hypothetical study outlined here, is discussed in Chapter 12.

When matching is a practical possibility and when regression artifacts are evaluated, we can feel somewhat more confident of conclusions from our results. But we should remember that what we have is still only a correlation, albeit a very carefully controlled one. Matching is sometimes useful, but it is not a cure-all. In our earlier example comparing schizophrenic subjects with others on mental test performance, if the schizophrenics still performed worse than the new matched control group, could we then conclude that schizophrenia *produced* inferiority in language usage? No, we could not. It could still be something else, some other difference between the two groups. We can never be absolutely sure we have matched on the relevant variables.

The study of experimental design is complex. In most chapters, we include a feature, From Problem to Experiment, that tells how to turn some issue or question into an actual experiment. We describe this feature next.

FROM PROBLEM TO EXPERIMENT

THE NUTS AND BOLTS

Problem *Conducting an Experiment*

Many of the decisions that go into creating an experiment are not clearly explained in journal reports of research. Although some of this brevity can be attributed to the economy imposed by journal editors who like short articles, a larger part is based on the assumption that experimental psychologists, or indeed psychologists researching any specialty, share a common background knowledge. This is true in all branches of science. For example, a physicist writing in a journal assumes that the readers already know that a dyne is a unit of force and will not bother to explain that term. Similarly, psychologists usually assume the reader knows what the terms *stimulus* and *response* mean, although these may be defined anyway. One purpose of this text is to give you some of the vocabulary necessary if you wish to read or write about psychological research.

Another problem for the new researcher is related to the "lore of the laboratory." "Everybody" knows there are certain "obvious" ways to perform certain kinds of research. These ways differ from area to area but are well known within each category. They are so well known that researchers seldom bother to explain them and indeed are quite surprised when new researchers are ignorant of these "obvious" tricks and techniques. Animal researchers often deprive animals of food for several hours before the experiment or keep their pigeons at a certain percentage of the weight the pigeons would attain if they had food continuously available. Although the reasons for this are obvious to the researcher, they may not be obvious to you. How does an experimenter know how many items to use in a memory experiment? How long should an experiment take? Why is one dependent variable selected from a set of what appear to be equally valid dependent variables? How many subjects should be used in an experiment? The From Problem to Experiment sections in the chapters of Part Two will answer such "obvious" questions as these.

From Problem to Experiment

All research aims at solving a problem. This problem can be abstract and theoretical or concrete and applied. The problem may arise from an observation made more or less casually, such as that people seem to be more aggressive during the summer. Here, the problem can be stated as "Why does summer heat cause aggression?" or even more skeptically as "Does high temperature cause aggression?" A problem may arise from an accidental discovery in a laboratory, such as the finding of mold on a piece of bread. Solving this problem—why is the mold growing here?—led to the discovery of penicillin. Finally, a problem may arise directly from a theoretical model, for instance, when we ask, "Why does reinforcement increase the probability of the occurrence of the behavior that preceded it?"

The first step the experimenter must take is to translate the problem into a testable hypothesis. The hypothesis then must be transformed into an experiment with independent, dependent, and control variables.

From Problem to Hypothesis A problem is, more or less, a vague statement that must be verified or a question that must be answered. Unless either is made specific and precise, it cannot be experimentally tested. Any hypothesis is a particular prediction, derived from a problem, often stated in this form: If A, then B. The crucial distinction between a problem and a hypothesis is that a hypothesis is directly testable, whereas a problem is not. An experimental test must be capable of disproving a hypothesis.

The purpose of any experiment is to test hypotheses about the effects of an independent variable(s) on the dependent variable. To do this, we must collect **data.** Once obtained, these data must be analyzed. Once analyzed, data must be reported. We briefly discuss these aspects in turn.

DATA

Obtaining Data

Outlining an experimental design does not establish all the conditions needed for data acquisition. Although the design tells you how to assign subjects to experiments, it does not tell you how to get the subjects. Without subjects, there are no data.

Psychologists who investigate animal behavior have much more control over subject selection than those who study humans. Although animal psychologists must bear the additional expense of obtaining housing and feeding their subjects, they can select the strain they wish to purchase and always have subjects available, barring some catastrophe.

Research with humans most often uses as subjects college students enrolled in introductory psychology. Provided that this participation is used as a learning experience for the student, it is considered ethical and proper (American Psychological Association [APA], 1987). If the experiment is not used as a learning experience, the experimenter should pay subjects. Since college students are a select population, experimenters need to be careful about generalizing results to other subject populations. For example, techniques from a programmed learning system designed to teach inorganic chemistry might not prove successful in the teaching of plumbing.

Random selection means that any member of a population has an equal chance of being selected as a participant. Furthermore, each selection is independent of other selections, so choosing one person does not affect the chances of selecting anyone else. Sometimes in a typical psychology experiment it can be difficult to specify the population being sampled (Gigerenzer, 1993). Even if subjects can be drawn randomly, exactly what population does a university subject pool represent? It is not even clear if the population of students taking required psychology courses are representative of all university students. Since the student population is now so diverse, representing people with many different ages and backgrounds, researchers need be careful about extrapolating results from the test sample to other populations.

Random assignment means that each participant in the experiment is randomly assigned to experimental treatments (Holland, 1993). This is a prudent technique

because it increases our ability to make causal inferences from the experimental results. Statistical implications of sampling are discussed in Appendix B.

After your sample has been selected and your design is fixed, one major decision remains. Should you test your participants one at a time or in a group? Both procedures have advantages and disadvantages. The biggest advantage of group testing is economy. It takes only 1 hour to test 30 participants for an hour as a group, whereas it takes 30 hours to test them singly. So, all other things being equal, it is faster, and therefore better, to test participants in groups. But there are many instances where all other things are far from equal. For example, take a listening experiment in which separate words are presented to left and right ears. One hurried doctoral student decided to save time and test her participants in a group. She forgot that unless participants were positioned exactly between the two loudspeakers, one message would reach one ear before the other message reached the other ear. This invalidated the independent variable. Of course, it would have been fine to test participants in a group if each person wore earphones, thus avoiding this difficulty. The other problem in group testing is the possibility that participants will influence one another, thus influencing the data. Perhaps a participant may cheat and copy answers from another, or the sexual composition of the group may alter motivation. Sometimes these problems can be prevented by placing participants in individual booths that prohibit social interaction.

Analyzing Data

The immediate result of an experiment is a large series of numbers that represent behavior under different conditions. As Sidman (1960) humorously describes it, scientists believe that all data are tainted at birth. Data belong to Chance or to Science—but never to both. Before the psychologist can be sure that data belong to Science, the demon Chance must be exorcised. This is done by a ritual called *inferential statistical analysis.*

Once statistical analysis tells you which data are reliable (did not occur by chance), you still have to decide which data are important. No mathematical calculation can tell what hypotheses are being tested, what is predicted by the theories, and so on. Statistics are never a substitute for thought. Statistical analysis is a theoretically neutral procedure that serves theory and hypothesis testing. Except in the case of a what-if experiment, the theories and hypotheses precede the statistics.

Because it is virtually impossible to grasp the meaning of the large set of numbers an experiment produces, data are usually condensed by descriptive statistics. The most common are the mean and the standard deviation. As part of the data analysis, means are calculated for each level of each independent variable, as well as for combinations of independent variables to show interactions.

Reporting Data

Data are presented in tables or figures. Figures are usually easier to understand. Figure 3.2 is a typical example of how results of an experiment are reported. The dependent variable is plotted on the **ordinate**—the vertical scale. The independent variable is graphed on the **abscissa**—the horizontal scale. More than one independent

variable can be shown in the same graph by using solid and dotted lines and/or differently shaped symbols for each independent variable.

Raw (unanalyzed) data are hardly ever reported. Instead, some descriptive statistic, such as the mean, is used to summarize data. Other statistics often accompany data to tell the reader about the reliability of these data.

Many different styles and formats can be used to report data. We recommend the format given in the *Publication Manual of the American Psychological Association,* which has become the standard reference in psychology and many other fields in social science. This book will tell you more than you would like to know about every aspect of preparing the report of an experiment. If it is not in the library or bookstore, you can purchase it through the Order Department, American Psychological Association, P.O. Box 2710, Hyattsville, Maryland 20784.

SUMMARY

1. An experiment is a controlled procedure for investigating the effects of one or more independent variables on one or more dependent variables. The *independent variable* is manipulated by the experimenter, whereas the *dependent variable* is observed and recorded. Experiments offer the investigator the best chance of eliminating or minimizing extraneous variation. Experiments are performed to test theories, to replicate and expand previous findings, or to show that prior research cannot be confirmed. Only rarely are experiments performed just to see what might happen.

2. Independent variables are chosen because an experimenter thinks they will control behavior. If they do not, this may mean that the manipulation was inadequate or that the experimenter was wrong. Dependent variables must be *stable*—that is, they must consistently produce the same results under the same conditions. *Ceiling* and *floor effects* result from an inadequate range for the dependent variable. *Control variables* are potential independent variables that are not manipulated during an experiment.

3. Most experiments test more than one independent variable at a time. In addition to providing economy, this allows the experimenter to gain important information about interactions. *Interactions* occur when the effects of one independent variable are not the same for different levels of another independent variable. Occasionally, experiments use more than one dependent variable.

4. Experimental design assigns subjects to different conditions in ways that are expected to minimize extraneous variation. In a *between-subjects design,* different groups of subjects experience different treatments. In a *within-subjects design,* the same subjects go through all treatments. The between-subjects design is safer, but the within-subjects design is more efficient. *Mixed designs* have some independent variables that are between-subjects and others that are within-subjects. In between-subjects designs, equivalent groups are formed by matching and by *randomization*. *General practice effects* and *differential carryover effects* in within-subjects designs are evaluated but not eliminated by counterbalancing. Control conditions provide a clear baseline against which the condition(s) of interest can be compared.

5. There are many pitfalls in experimental design. *Demand characteristics* result from the subject's knowledge that he or she is participating in an experiment. *Experimenter effects* are artifacts introduced accidentally, when the experimenter (through behavior or individual characteristics) provides clues regarding the purpose of the experiment or influences the subject systematically. Experimenter effects can be minimized by the use of machinery to preclude subtle differences in the experimenter's behavior.

6. Selecting participants from some population is called **sampling.** *Random sampling* means that each member of the population has an equal chance of being selected. It is more efficient to test subjects in groups, but care must be taken to avoid contaminating the experiment.

7. Quasi-experiments in psychology often employ subject variables. These variables are measures such as age, IQ, mental health, height, hair color, sex, and the myriad other characteristics that differ from one person to the next. Such variables are determined after the fact, since they are often inherited dispositions (or at least, people come to the psychological study with the variable already determined). Because it is not possible to assign people randomly to the conditions of interest, studies that use subject variables are inherently correlational in nature.
8. To attempt cause-and-effect statements from manipulation of subject variables, researchers often match subjects on other variables. Thus, if a researcher were interested in the effects of hair color on performance in some task or on the reaction from others in some situation, he or she would attempt to control as many other variables as possible to ensure that hair color was the only aspect on which people in the various conditions differed. Matching is often a useful tool for these purposes, but one must be certain that the possibility of regression artifacts does not cloud the conclusions.
9. *Regression to the mean* refers to the fact that when a subgroup with extreme scores is taken from a larger group and retested, members will tend to score nearer the mean of the whole group on the second test. If, in matching two groups on the basis of a first test, the researcher is taking high scorers from a group that generally does poorly and low scorers from a group that generally does well, then even if the groups are not treated differently in an experiment, the researcher can expect them to score differently on a second test—simply because of regression to the mean. This problem is referred to as a *regression artifact*.

KEY TERMS

abscissa
baseline
between-subjects design
ceiling effect
control condition
control group
control variable
counterbalancing
critical experiment
data
demand characteristics
dependent variable
differential carryover effects
double-blind experiment
experiment
experimenter effects
floor effect
general practice effects
Hawthorne effect
independent variable
interaction
level
matching
mixed design
null results
ordinate
quasi-experiments
random assignment
random selection
randomization
regression artifact
regression to the mean
sampling
small-*n* design
stability
subject variable
synergism
what-if experiment
within-subjects design