CHAPTER 1

AN INTRODUCTION TO EXPERIMENTS

WHY METHODOLOGY? WHY EXPERIMENTS?

"I'm not planning to do any psychological research myself, so why should I study research methods?" This is a question students ask us from time to time, and it deserves some answers.

The first answer is that you never know when you might be called upon to carry out some "research." No matter what field you enter, you may feel a need to improve your performance or that of your subordinates. You will inevitably run up against problems, and you may have to try out different ways of solving them. People in business are constantly striving to figure out ways to improve sales or productivity; doctors test new ways to persuade patients to follow their advice; lawyers try out new strategies for successful argument or negotiation. An understanding of the logic of research design is as useful in improving the informal research questions that continually arise in our professional and personal lives as it is in designing formal research studies.

The second answer is that understanding how to do research prepares us to *evaluate* the research we read about. In this information age, we are inundated with communications from the press and television media: If parents share custody of their children after divorce, the children will be better off than if only one parent has custody; coconut oil raises cholesterol and olive oil lowers it; bottling up one's emotions leads to illness; "back to basics" education results in higher S.A.T. scores; and so on and so on. As consumers of this information, we face the task of figuring out what claims to believe, of separating reasonable journalistic conclusions or interpretations from inferences that are biased or just plain wrong. An understanding of research methods prompts the consumer to ask "How do they know?" when faced with news of a new discovery and provides the tools to answer that question. Quite often it turns out that "they" don't know.

Finally, studying the methods of social psychological research can improve people's reasoning ability more generally. Darrin Lehman and his colleagues (Lehman, Lempert & Nisbett, 1988) found that training in social psychology improved students' ability to reason about everyday problems. Because the topics studied by social psychologists are quite similar to the problems we face in everyday life, learning to think about them the way researchers do creates skills that can be transferred to real-life problems.

"Why does your book place such a heavy emphasis on the *experiment*, when there are quite a few other methods for conducting psychological research?" This is a question our colleagues ask us from time to time, and it, too, deserves an answer.

The reason is that a true experiment is the best method for finding out whether one thing really *causes* another. Very often our questions are questions about causality. Does happiness *cause* people to behave more kindly? Does back-to-basics education *cause* higher S.A.T. scores? Does violent television programming *increase* aggression? Does enforced cooperation *reduce* racial prejudice? A well-conducted experiment can provide a definitive answer to questions like these, at least in its own particular context. Finding out whether the results apply in other contexts typically requires further research.

But the experiment is not the only method in the researcher's repertoire, and sometimes the student of human behavior chooses a different strategy. There are three common reasons for conducting nonexperimental research. The first is that the researcher is not interested in a question about cause and effect. A pollster, for example, may be interested in predicting who is going to win the next election, without caring *why* one candidate will be favored over another. Or, a researcher may simply want to demonstrate the *existence* of some phenomenon: that people from all over the world use the same facial expressions to communicate emotions or that people are prone to certain kinds of mistakes in reasoning—for example, they overestimate the "normality" of their own tastes and behavior (Ross, Greene & House, 1977), they can't resist coming up with logical explanations of events that are really random (Feller, 1968; Ward & Jenkins, 1965), they overestimate the likelihood of events that are vivid and more easily called to mind (Gregory, Burroughs & Ainslie, 1985; Sherman et al., 1985). For questions that do not involve causality, the experiment is not essential, although it may be very useful.

The second reason for conducting nonexperimental research is that in some situations an experiment would be unethical or impossible. If we want to know whether marriages of dissimilar people are more likely to end in divorce than marriages of similar people, we have to study the question nonexperimentally. We cannot force a hundred women to marry men who resemble them and a hundred others to marry men who are totally unlike them. There are a host of topics—mental illness, powerful emotions, the deterrent effect of the death penalty—that we are morally unwilling to impose upon people. There are a

host of others—gender, race, birth order—that we can't impose upon people. Because many extremely important questions involve qualities that cannot be controlled experimentally, researchers have devised increasingly sophisticated ways of conducting nonexperimental research to reveal causes. We will discuss some of these methods later on. They range from nearly worthless to fairly close approximations of true experiments, but they are all approximations. The better approximations *are* better because they are more fully informed by an understanding of the logic of the true experiment. Because the true experiment is the standard by which studies of cause and effect are evaluated, and because it is impossible to understand the virtues and shortcomings of the varieties of nonexperimental research without a thorough understanding of experimental research, we begin with experiments and emphasize them throughout this book.

The third reason that researchers do nonexperimental research is to *supplement* experimental research, so that the advantages of one kind of research can compensate for the shortcomings of the other: Two methods are better than one. Suppose you want to find out whether there is racism in criminal sentencing. You might first do an experiment, giving jury-eligible citizens a series of cases which differ only in that half the defendants are black and half are white. Thus, some subjects would hear about an armed robbery committed by a white man, and others would hear about the exact same events, only the perpetrator would be black. If the subjects recommended longer sentences for the black defendants, you could be sure that this was because they were black. This would be an enormously important finding, one very difficult to prove in any other manner. Nonetheless, your study would be open to criticism that your subjects were not *really* sentencing anybody and that their experience differed from that of real jurors in innumerable ways. So you might go out and collect data on hundreds of real trials, recording not only the race of the defendant but all the other factors that might lead to harsher sentencing—what the crime was, who the victim was, and so on. If blacks received longer sentences in this study, you wouldn't know for certain it was because they were black. It could be some other factor you failed to take into account—for example, the quality of the attorney. But you would know that your experimental results hold up in the real world. The two studies together are more persuasive than either alone: The first tells you that race influences sentences; the second tells you that real jurors behave like your subjects.

THE NATURE OF EXPERIMENTS

A scientific investigation starts with a question. Why do people yield to group pressure? Do people like something better if they have had to work hard to get it? How does a steady diet of violent television shows affect a person's behavior? Any question can be studied scientifically, provided that it involves something that can be observed. After formulating the question, the scientist must decide exactly what to observe in order to answer it. For many questions, the

scientist's next task will be to find a situation in which to observe the phenomenon. A researcher who is interested in the effects of different kinds of college curriculum on graduates' subsequent success, for example, will have to specify the types of curriculum that might make a difference. For some questions, the scientist may even have to wait for circumstances in which to make observations. Social psychologists who want to study people's responses to a natural disaster will have to wait for a flood, tornado, earthquake or other calamity to happen. Similarly, astronomers must wait for the earth to move into a particular position before they can make specific observations.

An experiment differs from other types of scientific investigation in that rather than searching for naturally occurring situations, the experimenter *creates* the conditions necessary for observation. There are several advantages to this procedure. First, by setting up the conditions, the experimenter has a better chance of capturing exactly what was intended. For example, Solomon Asch deliberately set up his experimental group so that the subject was faced with several other people who steadily and consistently disagreed with him about an apparently obvious judgment. It is hard to imagine a natural situation which so clearly and forcefully pits the evidence of one's senses against the evidence of one's peers.

Second, the experimenter can *control*—and thus systematically *vary*—conditions in order to study the same general situation with and without the crucial element. In studying the effects of group pressure, Asch arranged his experiment so that all subjects were members of a group in which the members' task was to judge the lengths of lines. In some groups the members called out their judgments so that everybody could hear them, and of course the subject had to let everybody hear his judgment, too. Since expressing his true opinion meant making a spectacle of himself by announcing the only deviant opinion in the group, there was a great deal of pressure toward conformity in these groups. In one group (the control group), the members simply wrote down their judgments, so the subject had no idea what the other members were saying and thus experienced no group pressure. If Asch had used a nonexperimental procedure, he would have had to try and find naturally occurring groups working on similar tasks in similar settings but differing in that some groups were characterized by high amounts of pressure on the individual and others were pressure-free. It is often extremely difficult, if not impossible, to find natural situations which are alike in all respects except the one that interests the experimenter.

Finally, and most important, the experimenter has the power to decide which individuals will be exposed to which conditions. Outside the experiment, people who are more independent may choose groups in which less pressure is exerted; less independent people may like to have their decisions made for them and so prefer groups with obvious pressure. Thus, if the individuals in the groups with the most pressure conformed more to their groups, this finding might not be due to the group pressure but to the fact that the people who chose those groups were more conforming to begin with. In an exper-

iment, the experimenter assigns subjects to treatments *at random*. Thus, whether a person happens to be a member of a group that exerts pressure is due to chance alone, and not to any prior differences among subjects. If the subjects in the group that exerts pressure make more errors in their judgments of the lines, the experimenter *knows* that this difference was caused by the group pressure and not by any prior differences in the individuals who were members of the two types of group. Because the experimenter can assign subjects to treatments at random, the experiment, unlike other procedures, can provide a strong basis for making statements about causality.

Experiments have certain general features in common. Because much of this book will be concerned with the actual procedures of experimentation, it is important to start out with a general overview of these procedures, introducing and defining the terms that we will be using in subsequent chapters. In order to make this presentation more concrete, we will turn again to the Asch experiment as an example of how an experiment is put together. Having looked at the Asch experiment through the eyes of the subject, we will now reexamine it from the point of view of the experimenter.

The experimenter begins with an idea, or a question, or a hypothesis. The idea may derive from a theory, from doubts about the validity of some previous experiment or theoretical formulation, from a concern with a social problem, or simply from curiosity about some kind of behavior. Asch was generally interested in the conditions of submission to group pressure, which he felt to be a problem involving important social issues. He was dissatisfied with earlier social psychological accounts, because they tended to give the phenomenon a label, such as "suggestion," which explained little, and because they paid too little attention to the processes that enable a person to *withstand* group pressure. Thus, like most experimenters, he was interested in the problem for a variety of reasons.

Independent and Dependent Variables

The next step is to translate a concern for a general problem area into a specific question. Asch's specific question was: What are "the social and personal conditions that induce individuals to resist or to yield to group pressures when the latter are perceived to be *contrary to fact*?" (Asch, 1965, p. 393), and he started out with one "social condition"—the presence of a unanimous majority. In many experiments the basic question is stated as a **hypothesis**, or prediction about the outcome of an experiment. Either way, an inquiry is made about a causal sequence. The antecedent event, or "cause," in the proposed sequence is called the **independent variable,** because the experimenter creates it and controls its variation; it is independent of all other causative influences. The experimenter is sometimes said to *manipulate* the independent variable. In the initial Asch experiment, the independent variable was group pressure—specifically, the pressure assumed to be exerted by a unanimous majority which gave false judgments. The "effect" in the causal sequence is called the

dependent variable, because the experimenter expects its value to depend on the changes introduced in the independent variable. In the Asch experiment the dependent variable was the number of errors the subject made in the direction of the false judgments of the majority.

Technically, a **variable** is any attribute which can assume different values among the members of a class of subjects or events but which has only one value for any given member of that class at any given time. Thus, height is a variable within the class of human beings (and other things); within the class as a whole height can assume a large range of values, but at any given moment a given human being can be only one height. Some variables, such as height, are *continuous* and can assume any value within some finite range. Others, such as virginity, are *discrete* and can assume only a limited number of values—in the case of virginity, only one of two: presence or absence. In a psychology experiment the independent variable is usually a stimulus event, and the dependent variable is a response made by the subject.

The values of the independent variable (or variables) that the experimenter chooses to use define the experimental **conditions.** At least two values of the independent variable are necessary in order to demonstrate that the variable is having an effect, since in an "experiment" with only one value of the independent variable, it is impossible to determine whether the results (the measurements of the dependent variable) have anything to do with the presence of the independent variable. In the Asch experiment there were two values of the independent variable: presence of group pressure (in the form of a unanimous majority) and absence of group pressure. In a study of this sort, in which there are only two conditions defined by presence *versus* absence of the independent variable, the subjects who receive the independent variable (those who had to make their judgments in the face of unanimous opposition) constitute the **experimental group,** and those who do not receive it constitute the **control group.**

Having formulated a question or hypothesis, the experimenter must decide how to turn it into a set of experimental procedures. One of the most important parts of this transformation involves translating the concepts of the hypothesis into specific, observable events. This substitution may be viewed as the **empirical realization** of the conceptual or abstract variables contained in the question. An abstract concept such as "group pressure" is "made real" in terms of the events actually experienced by the subjects—thus the term "empirical realization." Usually in social psychology the process of creating an empirical realization is intimately bound up with decisions about the overall staging and context of the experiment. Thus, in the Asch experiment the impact of the group pressure depended on a great many elements of the situation—the apparent simplicity of the task, the public announcements of the members' judgments, the unanimity of the judgments by the other "subjects," and the consistency of those judgments through time. Certainly, this was an extreme form of group pressure, unlikely to be encountered outside the laboratory, and Asch intended it to be so. In many real-life situations, people are able to avoid the dilemma created by a contradiction between their own beliefs and the norms of

a group. For one thing, the ambiguity of some situations allows people to avoid even perceiving that there is a dilemma. Also, the fact that people typically are not required to state their opinions publicly may allow them to assume that they are not the kind of people who yield to others. In addition, it is often possible to explain away discrepancies by referring to factors outside the situation or by assuming that the majority are motivated by a desire to persuade others or to play devil's advocate and don't "really mean" what they are saying. Asch wanted to remove these extraneous defenses in his empirical realization of the concept "group pressure" in order to study it in the pure case, forcing the subject to face the dilemma and to resolve it by yielding or by standing firm. Once having determined what happens in this extreme case, he could then make the situation less extreme and compare the results with those obtained in the initial experiment.

In Asch's experiment the empirical realization of the dependent variable followed naturally from the rest of the situation; the conceptual dependent variable, yielding, was simply realized as the number of times a subject went along with the group and gave the wrong answer.

For the kind of complex variables studied by social psychologists—variables such as guilt, anxiety, self-esteem, and group pressure—there is no one "right" empirical realization. Some are better than others, and in Chapters 7 and 8 we will discuss some of the characteristics of a good empirical realization and some of the techniques for creating one. More than one empirical realization may capture essential features of some variable such as "group pressure," but they may emphasize different features. An experimenter interested in studying how people evade facing the dilemma posed by conflicting group and personal values, for example, might create a situation in which the pressure was less obvious or in which the subject could attribute the group's disagreement to some external factor. This experimenter's data on yielding might look very different from those obtained by Asch. Such was the case in an experiment by Lee Ross, Günter Bierbrauer, and Susan Hoffman (1976), in which the experimenters manipulated possible *reasons* for the incorrect judgments of confederates. They found that when the subjects could find some *reason* for the difference between their perceptions and the majority answer—even if it was not a very convincing reason—they no longer yielded. Of course, their results did not invalidate Asch's findings about the effects of group pressure. When different empirical realizations of a concept produce different kinds of behavior, it is an indication that the original conceptual variable—in this case, group pressure—is too general and needs to be differentiated into a number of less general related variables. Ross and his colleagues showed that if people can convince themselves that the majority is responding to a "different situation," they will be much less likely to give in to group pressure. By this process of differentiation, social psychologists arrive at a more detailed and comprehensive understanding of a general category of behavior—such as people's susceptibility to group conformity pressures.

Sources of Error

Before actually beginning to run the experiment, the experimenter must also consider various types of *error* that might arise. Although we are interested in varying only the independent variable, in practice it is impossible to design an experiment in which nothing except the independent variable affects the outcome. This is true primarily because we are dealing with human beings. Anything besides the independent variable that affects subjects' behavior is a source of error. Subjects bring personal variables into the experiment—age, background, alertness, intelligence, and so on. Subjects may differ with respect to their experience with the type of independent variable being manipulated and with the dependent variable being measured. For example, the average number of errors made by a subject in the Asch experiment was 3.84 (out of 12 trials), but there was wide variability; some subjects never gave in to the group, and others displaced their estimates toward the majority in more than half the trials. Asch made a point of analyzing individual differences among subjects that might have affected their tendency to yield. One factor that often differentiated between the yielders and nonyielders was self-esteem, with the yielders showing "primary doubt and lack of self-confidence" (1965, p. 397). Variables such as these are called **subject variables;** since the experimenter can exert no direct control over them, they are typically sources of error.

Two types of error can affect the outcome of an experiment. The first type, called **random error,** refers to extraneous variables whose *average* influence on the outcome is the same in both (or all) conditions. Subject variables and minor events that occur during particular experimental sessions contribute to random error, as do all the other extraneous influences on the subject's behavior which are not controlled by the experimenter and which are equally likely to occur in any of the conditions of the experiment. In this sense, random error constitutes the "noise" in the system, from which the experimenter is trying to extract a meaningful "signal" or *consistent* type of variation produced by the treatments administered.

If a given source of random error tends to raise the subjects' scores in one condition, it will also raise the scores of those in the other conditions. This will decrease our confidence in the actual numerical levels of the scores (since they will all be artificially inflated), but it will not decrease our confidence in the *differences* between the two conditions, since both sets of scores are raised by the same amount.

If, as is more usual, a given source of random error simply increases the range of scores in both groups, it is possible that this increased variability will obscure the effects of the independent variable, and the experimenter may erroneously conclude that the treatment had no effect. Had the investigator been able either to eliminate more of the random error (reduce the noise) or to use a stronger treatment (generate a stronger signal), the treatment difference might have been large enough to show up against the noisy background.

The second type of error, called **systematic error** (constant error), is much

more dangerous. Whereas random error typically increases the baseline variability in both conditions, systematic error tends to influence all the scores in one condition in the same direction and to have no effect, or a different effect, on the scores in the other condition. Thus, systematic error can affect the size of the difference between the two conditions, thereby distorting the experimenter's source of information about the effects of the independent variable and possibly vitiating the results of the whole experiment.

Obviously, it is in the experimenter's best interest to attempt to reduce both kinds of error, but of the two types, it is more important to eliminate systematic error. If there is too much random error in an experiment, a true relationship between the independent and dependent variables can be obscured, so that the experimenter might erroneously conclude that the variables are unrelated. *Systematic* error, however, can make it look as though two variables are related, when in fact they are not; thus, the experimenter concludes that the hypothesis has been confirmed when it has not. In the first case, the experiment will not be published, and the worst that can happen is that the experimenter will fail to discover an interesting phenomenon which a more carefully controlled experiment would have revealed. In many cases, the experimenter is interested enough in the hypothesis to attempt another experiment on the same topic, improving the procedure, so that the true relationship does not remain buried for long. In the case of systematic error, however, the experimenter often publishes the spurious finding, and it remains in the literature, influencing experimental work for a long time and stimulating efforts that could more productively be directed elsewhere. Unfortunately, social psychology is far from being an exact science, and once such a spurious finding becomes part of the literature, it can be very difficult to disprove.

One way of eliminating systematic error is to convert it to random error. Different sources of variability, such as subject variables, the experimenter's behavior, time of day and distractions and extraneous events within the experimental setting, are not in *themselves* automatically random or systematic influences. Whether they affect the subject's responses randomly or systematically depends on whether certain experimental precautions have been taken. For example, suppose that some extremely nearsighted subjects in the Asch experiment were unable to see the lines well enough to judge them accurately. If for some reason all these subjects were placed in the experimental group, their inability to judge the lines accurately would spuriously increase the number of errors made by subjects in that group and perhaps cause the experimenter to draw unwarranted conclusions about the effects of group pressure. Perhaps the experimenter unintentionally biased the data by assigning all the nearsighted subjects to the experimental group; perhaps they were all friends and came to the experiment together, and the experimenter happened to be running only the experimental group that afternoon. Whatever the reason, anything that results in the overrepresentation of subjects of a given type in one experimental condition is a potential source of systematic error. If the experimenter had assigned each subject to the experimental or control group *at random,* however, the error would be *random* error. It would be highly unlikely that all nearsighted subjects would end up in the experimental group, and thus the increased misjudgments made by these subjects would be balanced out by similarly inflated misjudgments made by their counterparts in the control group.

Random Assignment of Subjects to Treatments

In general, **random assignment** is one of the experimenter's most important tools for ruling out the dangers of systematic error. It is so important, in fact, that it is considered the criterial attribute for defining a study as an *experiment.* The most common variety of random assignment is the random assignment of subjects to experimental conditions, as in the example of the nearsighted subjects mentioned above. The experimenter's goal is to make sure that none of the myriad extraneous factors which might affect a subject's behavior in the experiment—such as nearsightedness, intelligence, or a bad mood—is more likely to be associated with one of the experimental conditions than with the other (or others). Since the conditions are defined by the independent-variable treatments, they must be kept free of any extraneous factors that may cause differences between them in order for the experimenter to conclude that the differences were *caused* by the independent variable alone. If subjects in one condition were more nearsighted, or intelligent, or unhappy, the experimenter could not conclude that the independent variable was affecting their behavior; any differences that show up in the behavior of the subjects in the two conditions might be due to differences in the type of subject assigned to the conditions in the first place.

Assigning subjects to conditions at random means that each subject who walks in the door has an equal chance of being placed in any one of the experimental conditions. If there are two conditions, each subject has a 50:50 chance of being in either one.[1] Thus, nearsighted subjects have a 50:50 chance of being in the experimental group or the control group, intelligent sub-

[1] With two conditions, one method for random assignment is simply to flip a coin, assigning "heads" to the experimental group and "tails" to the control group. A table of random numbers is a better and safer tool for random assignment and has the advantage of being adaptable to any number of conditions. A table of random numbers is simply a table in which the digits 0 to 9 are arranged in random order in a long sequence. The experimenter decides arbitrarily that each experimental group will correspond to a number or type of number. With two groups, you might decide that the odd numbers will represent the experimental group and the even numbers the control group. Then you pick a place on the table to begin, again arbitrarily. If the first number you come to is odd, the first subject is assigned to the experimental group. You continue through the table in order until you have run as many subjects as you want. With three conditions, you might decide that the numbers 1 to 3 will correspond to the first condition, 4 to 6 to the second, and 7 to 9 to the third (ignoring zero). If it is important to come out with an equal number of subjects in all groups, a table of random permutations is used. In this kind of table, the numbers 0 to 9, for example, are put in random order once, so that all ten digits appear. Then the next ten digits are 0 to 9 again, in some other random order. This assures that after every ten subjects you will have equal numbers in the experimental and the control group.

jects have a 50:50 chance, unhappy subjects have a 50:50 chance, and so on for *every* kind of characteristic that can differentiate among subjects. The end result is that there will be roughly equal numbers of nearsighted subjects in the experimental and control groups, likewise for intelligent subjects, unhappy subjects, and so on for *all* kinds of traits. By randomly assigning subjects to conditions, the experimenter can be sure that *no* subject variable is more likely to occur in one condition than in the other and thus that no subject variable is a source of systematic error.

We can look at random assignment in another way—as the "great equalizer." Because random assignment ensures that all extraneous factors that might influence the subject's behavior in the experiment are approximately equal in the two (or more) conditions, we would expect that if we *left out* the experimental treatments and ran both groups of subjects as control groups, the average scores of these two groups on the dependent-variable measure would be the same. Not every subject would have the same score, of course. Even in Asch's control group, in which there was no group pressure, 5 percent of the subjects made errors of judgment. With random assignment, we can assume that in a group of subjects exposed to group pressure, about 5 percent of them would have made mistakes even without the pressure. Since in the Asch experiment 74 percent of the subjects who were exposed to group pressure made at least one mistake, we can conclude that the independent variable had a large effect.

Factors other than subject variables can cause systematic error, if they are associated with some conditions more than with others, and these factors may also be eliminated by random assignment. For example, if there is more than one experimenter, it is important that each experimenter run about the same number of subjects in each condition. If one experimenter ran subjects only (or mostly) in the experimental group, and the other experimenter ran subjects mostly in the control group, differences between these two conditions could be due to differences in the personalities or techniques of the two experimenters. Thus, subjects should be assigned to experimenters at random.

Holding Variables Constant

Although random assignment is essential for eliminating systematic error, it cannot reduce the amount of random error, or "background noise," in the experiment. If the treatment is only one of a large number of factors influencing the subject's behavior in important ways, its influence may not be strong enough to stand out above the variability introduced by all the other extraneous factors. A common means of controlling random error is to *hold* important extraneous variable *constant* at a single level (or to reduce the possible levels to a more limited range). In his initial experiment Asch always used the same sets of lines (holding stimulus materials constant) and always used a unanimous majority of seven accomplices (holding size of group constant). In regard

to our hypothetical example of the nearsighted subjects, Asch might also have given all subjects eye tests before they participated in the experiment and eliminated those who couldn't see well, thus holding visual acuity constant at 20-20. The principle behind the technique of holding a variable constant is that the less an extraneous variable is allowed to vary, the less it can affect the dependent variable. Of course, it is never possible to control all sources of random error; the experimenter must use judgment in deciding which extraneous factors are most likely to produce large fluctuations in the particular dependent variable being measured.

In all social psychological experiments some factors are held constant, and others, typically considered less important, are allowed to vary at random. Although holding variables constant may appeal to our sense of tidiness, it does have some drawbacks in that it places limits on the conclusions that can be drawn from the experiment. By preventing a variable from affecting the subjects' behavior in a given experimental setting—by holding it constant—we give up the possibility of finding out whether that variable would have affected subjects' behavior had it been allowed to vary. Sometimes the experimenter doesn't care much whether it would have had an effect, because it is not relevant to the particular enquiry. Often, the investigator *knows* it would have had an effect but wants to eliminate this effect in order to observe the effect of some other variable against a clear background. For example, it is frequently found that women are more likely to receive help than are men, especially the kind of "chivalrous" assistance elicited in experimental studies of helping behavior. In addition, men are more likely to help women than they are to help other men, although women are as likely to come to men's aid as to women's (Eagly & Crowley, 1986). Now suppose an experimenter wants to study something else that might influence people's willingness to help others—for example, to find out whether feeling happy makes people more helpful. An experimenter who runs subjects of both sexes—or who uses both male and female confederates—can expect to find very different levels of helping, depending on the gender combination of needy person and helper, and this complicated "background" may make it very difficult to tease out the effect of a happy mood. Therefore, in conducting experiments on helping behavior, investigators often hold the sex of the subject and/or sex of the "needy" confederate constant—for example, using only male subjects and male sufferers.

Although this procedure is valuable, holding a variable constant at some level restricts the generality of the experimenter's conclusions to situations involving that particular level. The helping researcher only knows that happy men are more likely to help other men; whether they are more likely to help women, and whether women's moods affect their helpfulness, remain unanswered questions. After Asch's initial experiment, it was clear that subjects tended to conform to the judgments of a unanimous majority of seven. But what about a unanimous majority of six? Or five? Or three? We might be tempted to make an educated guess, but conclusions from the actual experi-

ment performed would have to be limited to majorities of seven. We have no information about the importance of the size of the majority in inducing conformity, and thus we cannot generalize to majorities of other sizes.

Now, suppose that Asch had not bothered to hold the size of the majority constant but had let it vary, say, between three and eight. If group size were *not* an important variable affecting conformity, he would have obtained basically the same results as he did with the constant majority of seven. As a result of his experiment, he could have generalized the conformity effect to a wider range of majority sizes. If group size *were* an important variable, however, some of his majorities would have induced subjects to make many misjudgments, and other majorities might not have induced subjects to conform at all. Thus, many *more* of the subjects in the experimental group would have behaved like members of the control group, making no errors at all, and the variability *within* the experimental group would have been increased. There would have been a large increase in random error, and it is quite possible that the overall effect of group pressure would have gone undetected. Allowing a potentially important factor to vary at random is a gamble. If it "pays off"—if the experiment produces significant results—the experimenter is more confident of the generality of these results than if the factor had been held constant. If the experiment does not show any difference between the groups, background error may have obscured a true difference that would have been apparent had the experimenter held the variable constant. Holding variables constant, then, is the safer procedure, although allowing them to vary at random is *potentially* more informative.

Systematic Variation

Besides holding a variable constant and allowing it to vary at random, the researcher has a third choice—varying it *systematically*, adding it to the experiment as a second independent variable. In effect, the experimenter asks a direct question about the additional variable instead of hoping to pick up some relevant information indirectly, as happens when the variable is allowed to vary randomly. In the Asch experiment, this would be equivalent to having several different groups of subjects who were exposed to group pressure, some subjects being assigned to a condition in which the majority consisted of seven people, some to a majority of six people, some to a majority of eight people, and so on. By comparing the number of errors made by subjects in each of these conditions, the experimenter gains a much surer understanding of the relation between size of majority and conformity than is possible by with either of the other methods.

In fact, Asch did vary the size of the majority systematically in a later experiment. Some subjects were members of pairs—they were faced with only one other person who gave wrong answers. Other subjects had to face a majority of two; others, majorities of three, four, or eight; and finally, some subjects were members of very large groups in which the size of the majority

ranged from ten to fifteen. Asch found that subjects could hold their own against a single other person who disagreed with them; no subject made more than one mistake, and most made none at all. When the majority consisted of two, conformity began to show up, subjects averaging 1.53 errors. With three people in the majority, the effect increased to its full strength, with subjects making an average of four errors. Larger majorities, even those of ten to fifteen, did not produce effects stronger that did a majority of three.

Interactions

Another advantage of varying more than one independent variable at a time is that this technique can provide information about how one independent variable *interacts* with the other. An **interaction** is a situation in which the independent variable has *different* effects, depending on the value of some other variable. If the other variable is also experimentally varied, the data will show the interaction. But if the other variable is either held constant or allowed to vary at random, information about the interaction may be suppressed.

To take a hypothetical example, suppose that Asch had run an experiment in which there were two independent variables—majority size and similarity of the people in the majority to the subject. To simplify things, we will imagine that only two levels of each of these variables are used; thus, the experimenter has (1) majorities of two and (2) majorities of four, and these majorities are made up either of people very similar to the subject or of people very dissimilar to the subject. Each subject is randomly assigned to one of these four types of groups.

Given this design, we can envision several possible patterns of results. If the size of the majority has a large influence on the number of errors the subject makes but the similarity of the group makes no difference, the results would look something like those shown in Figure 1-1. This graph illustrates a *main effect* for majority size and *no effect* for the similarity variable. (A main effect is an effect caused by a single variable.) In other words, large majorities (defined here as four people) cause people to conform more than do small majorities (defined here as two people), and they have this effect whether or not the members of the majority are similar to the subject.

We can also envision a situation in which people tend to conform to the judgments of those who are similar to them but are able to maintain their independence when they perceive the others as dissimilar, regardless of how many others there are. In other words, we can imagine a situation in which there is a *main effect* for similarity, with majority size having no effect. This situation is illustrated in Figure 1-2.

Or, we can envision a situation in which both the size of the majority and its similarity to the subject make a difference. For example, subjects may tend to agree with similar people more than with dissimilar people, and they *also* tend to agree with large majorities (four people). There are then *two main effects*, one for similarity and one for majority size, as shown in Figure 1-3.

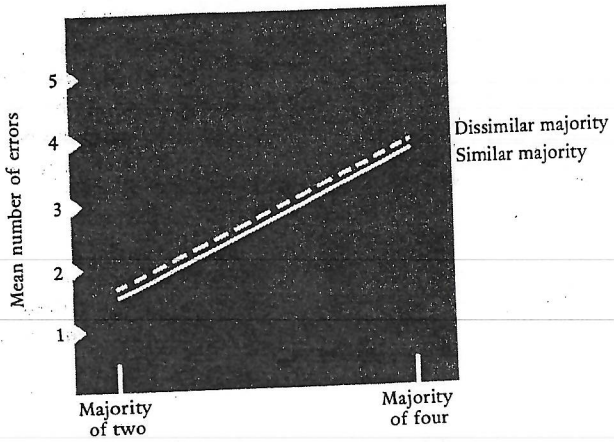

FIGURE 1-1
Main effect for group size; no effect for similarity.

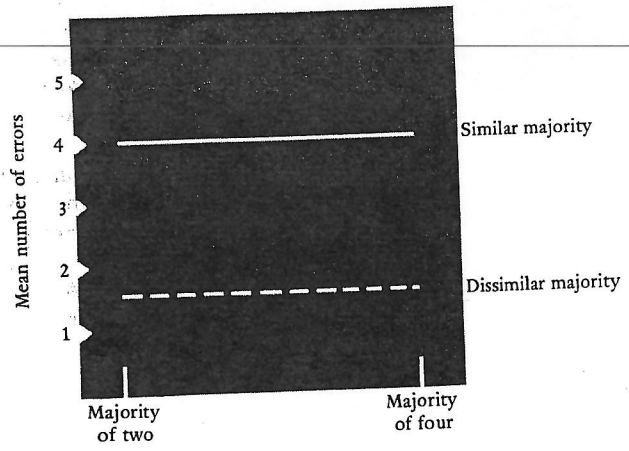

FIGURE 1-2
Main effect for similarity; no effect for group size.

In each of the three figures, the two lines have the same slope. This means that each variable has a constant effect, independent of the effect of the other variable. In Figure 1-3, the effect of increasing the majority size from two to four is to increase the number of errors by an average of two, regardless of whether the group is made up of people who are similar or dissimilar to the subject. The effect of making the group similar instead of dissimilar is to raise

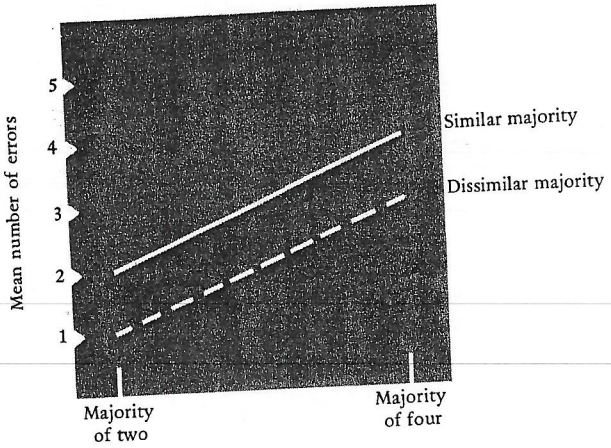

FIGURE 1-3
Main effect for group size; main effect for similarity.

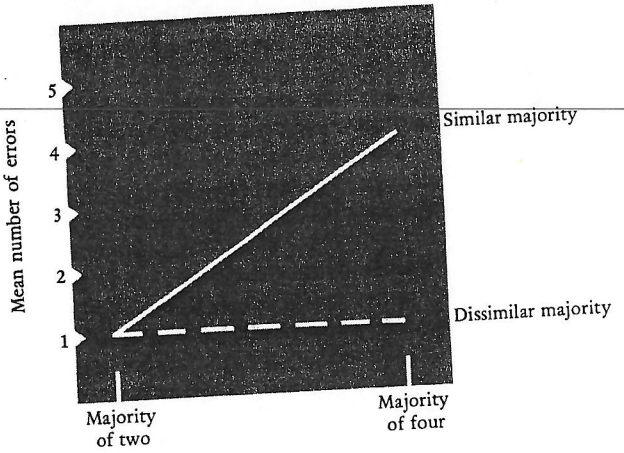

FIGURE 1-4
An example of an interaction between majority size and similarity.

the number of errors by an average of one, regardless of whether the majority size is two or four.

In an *interaction*, the slopes of the two lines will differ. This means that the effect of one independent variable (or both) is not constant, but varies depending on the value of the other independent variable. In Figure 1-4 we see one type of interaction that might exist in our hypothetical example. Majority size

has an effect, but only when the group is similar; the effect of majority size depends on the value of the other independent variable, similarity of the group. It takes a *large* group of *similar* people to get the subject to make many mistakes. Neither small groups nor dissimilar groups have any effect. By varying two variables at once, their combined effects are readily apparent.

What would have happened if the experimenter had held one of the variables constant? Obviously, part of the picture would be missing. For example, if similarity had been held constant—if only similar groups had been used—the experimenter would have found a strong effect and might therefore be tempted to overgeneralize, lacking the evidence of the effects of majority size in *dissimilar* groups which is available in Figure 1-4. In effect, the experimenter would not know whether the effects of majority size were those of Figure 1-1, Figure 1-3, or Figure 1-4, having only the data for similar groups, and these data are the same in all three situations. However, if only *dissimilar* groups had been used, the experimenter might have erroneously concluded that majority size does not affect conformity.

The experimenter would also obtain an incomplete picture of the situation by holding majority size constant, varying only similarity. Holding majority size constant at four, for example, the experimenter might assume that similar groups always tend to produce more conformity; at two, the experimenter might give up the whole line of research, since the results would indicate that similarity makes no difference.

What would have happened if similarity had been allowed to vary at random? Since the groups that were more similar would tend to influence the subject strongly and those that were less similar would have only a weak influence, we would expect to get an effect somewhere in between the strong effect for all-similar groups and the zero effect for all-dissimilar groups. Figure 1-5 shows what the data might look like.

Whether or not the experimenter decides that the data conform to a particular hypothesis about the effect of majority size, it is clear that less information is available than if similarity had also been varied. An experimenter who examines the data closely may note that the effect of majority size is not very uniform, as it would be were there a single main effect. In the large-majority condition, some subjects conform a great deal and some not at all, and the experimenter might be led to conduct further experiments, introducing other variables, to find out why. However, if the diluted effect of group size does not attain significance, the investigator might decide that the hypothesis was no good and therefore abandon the study of majority effects.

Artifacts

Although there are inherent difficulties in both strategies—holding variables constant and allowing them to vary at random—both are necessary techniques in any social psychological experiment, since there are so many possible variables that affect social behavior that it is impossible to vary them all systematically in a single experiment. Some random error, as well as many limitations on the generality of the findings, are inevitable. It is crucial, however, to make sure that no other variables change *with* the independent variable, always taking on one range of values in the experimental group and another in the control group. Such variables can be sources of systematic error in that they can cause differences between the two groups, differences which are mistakenly attributed to the independent variable. An extraneous variable that varies along with the treatment is said to be *confounded* with the independent variable. When the extraneous variable is a relatively trivial methodological event, the error is often called an **artifact.**

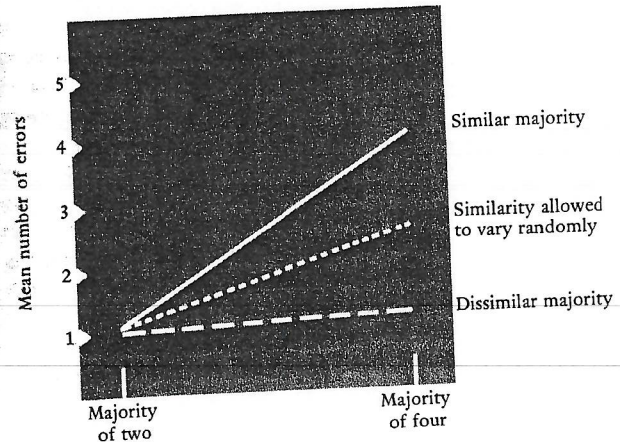

FIGURE 1-5
Results of a hypothetical conformity experiment, with similarity allowed to vary at random.

To take a hypothetical example, if the subjects in the experimental group undergo a complex and interesting experience during the presentation of the independent variable, but subjects in the control group simply come into the lab and sit and wait for the same amount of time, differences between the two groups could be caused by the boredom of the control subjects rather than by the effects of the independent variable on the experimental subjects. If in the Asch experiment the lines to be judged were flashed on a screen and removed before the judging started, subjects in the control group would be writing down their judgments immediately after the stimulus disappeared, but subjects in the experimental group would have to wait until all the "subjects" before them had given their answers. This longer delay between seeing the lines and giving their judgments might have caused the subjects to forget which line was really longest and thus could have operated to increase the number of errors made, even in the absence of group pressure.

Two special kinds of artifact are **demand characteristics** (changes in the sub-

jects' behavior simply from the experience of being in an experiment) and **experimenter bias** (changes in the subjects' behavior resulting from subtle hints unintentionally given off by the experimenter which let the subjects know how they are expected to respond). These two types of systematic error will be treated in detail in Chapter 9.

Alternative Explanations

Sometimes, the whole experimental procedure may allow for two substantively different and theoretically interesting interpretations of the results. In other words, the variables are not well enough defined so that the expected results *necessarily* indicate the validity of the experimenter's hypothesis. Someone reading the experiment may think of an *alternative explanation* which also fits all the experimental data.

A tradition of research on the *risky shift* illustrates what we mean by the term "alternative explanations." In the 1950s, many people believed that groups were more cautious than individuals in making decisions; that is, members of a group would be more level-headed and would tend to avoid extreme behavior. James Stoner (1961) set out to test that proposition. Briefly, he had people read about fictitious characters who had to make a choice between two alternatives: to take a risk or to adopt a more conservative stance. Participants were asked to decide how much risk they would advise the character to take. After they made their decisions individually, participants met in groups to discuss the situations and to reach agreement on what the group would advise. Were group decisions after discussion more cautious than the individuals' earlier decisions? Surprisingly not. By and large, groups recommended that the characters adopt a riskier strategy than did the average individuals, hence the term "risky shift."

These unexpected results initiated a whole line of research on group decision making in many different countries and with people of all ages from all walks of life. Investigators using Stoner's choice dilemma found consistent evidence of the risky shift. It seemed to be an amazingly robust finding. Eventually, however, some researchers uncovered the existence of a *conservative shift*. That is, under some circumstances, group decisions are more conservative than the average individual decision. Given these conflicting findings, it became the task of investigators to explain them. Are groups *really* more willing to take risks than are individuals?

Careful study revealed that group decisions are not always more risky than individual decisions; neither are they always more conservative. There is one reliable characteristic of group decisions, however: They tend to be more *extreme* than the average individual decision. That is, if individuals initially tend to favor conservative solutions to choice dilemmas, they are likely to shift to an even more conservative position when they get together as a group. Similarly, if individuals tend to favor risky solutions, group decisions are likely to shift to a more risky position. Thus, the risky shift demonstrated by Stoner and

the conservative shift demonstrated by others were both examples of a more general tendency of group discussions to exaggerate individual group members' initial inclinations. This more general phenomenon has been labeled *group polarization*.

The preceding story serves as an example of a "reasonable" explanation for research findings and an even *more* reasonable alternative explanation. The story also has a moral: Researchers should try to anticipate alternative explanations and design experiments that will rule them out by adding extra control groups, by changing the design, or, if these are not feasible, by collecting additional data which may help to discriminate between two plausible explanations. For example, had Stoner and others used two kinds of choice dilemmas—yielding both conservative and risky decisions by individuals prior to group discussion—or had investigators used different measures of risk taking, this interesting group phenomenon might have been understood sooner.

If some potentially interesting variables have been held constant or allowed to vary at random, the experimenter often decides to run a *replication* of the experiment, this time systematically varying one or more of these variables to find out what effect it has within the context of the experimental situation. Or, if an alternative explanation has been suggested, the experimenter may run a replication in which the conditions are changed in such a way as to rule out the alternative explanation. A **replication** is nothing more than a repetition of the experiment. In a **direct replication** the experimenter tries to make the conditions exactly the same as they were in the original experiment in order to see if the experimental effect is a stable (or reliable) one. If the replication is successful—if the results in the two experiments are the same—the experimenter will be more confident that the observed effect is a stable one that will be consistently achieved in the same conditions. We hasten to add, however, that direct replications are relatively rare in social psychological research. More often, investigators replicate and *extend* previous research; that is, direct replications are typically only part of a study in which additional variables are studied as well.

In a **systematic replication** the experimenter *varies* some quality of the original situation in order to resolve ambiguities or to add new information about the variables controlling the subjects' behavior. Asch's follow-up experiment with different majority sizes may be viewed as a systematic replication of the original experiment. It provided additional information about the relationship between the size of the erroneous majority and its effectiveness in inducing the subject to conform.

NONEXPERIMENTAL METHODS OF RESEARCH

So far, we have been discussing research that falls into the category of the *experiment*, a study in which the investigator has some control over the independent variables and can assign subjects to conditions at random. By *manipulating some variable*, the experimenter creates differences in the experiences of

two or more groups of subjects; and by *assigning subjects to the groups at random,* the experimenter creates groups which are equivalent in all respects except their experience with the independent variable. The experimenter can be reasonably sure that any differences in the behavior of the subjects in the groups are due to the differences in the treatments, since the random assignment of subjects to conditions has ruled out everything that might have made the two groups different to begin with. If the experiment has been conducted properly, the investigator can legitimately say that the treatments *caused* differences in behavior. Thus, Asch was justified in claiming that in his experiment, unanimous majorities *caused* many people to distort their judgment so as to correspond more closely to the majority judgment.

Only in an experiment can we *control the independent variable* and *assign subjects randomly* to conditions. For this reason, the controlled experiment is the only research method that allows us to make unequivocal cause-effect statements. The experiment is thus an extremely important method of conducting empirical research, but it is by no means the only method. It is true that in other kinds of research we invariably sacrifice the chance to make definite statements about causality, but below this rigorous standard extends a long continuum of methods, ranging from those that are simply suggestive, providing ideas in the same way as a novel or a newspaper article might, to designs that may provide a tenable scientific basis for inferring causality, designs hardly distinguishable from experiments.

Correlational Studies

Many important and interesting questions are not amenable to experimental research, because the experimenter cannot control the presumed antecedents or cannot assign subjects at random. Instead of introducing treatments, the researcher makes observations of events as they occur in nature. This does not mean that the research is "unscientific"; anyone who lacks faith in the development of a science through the exclusive use of correlational methods need only look at the history of astronomy. It does not even mean that statements of cause and effect are forever destined to be regarded as questionable. We do not raise our eyebrows and ask for experimental evidence when we read that eclipses of the moon are caused by the earth passing between the moon and the sun or that the odd footprints scattered across some outcroppings of Connecticut red sandstone were made by dinosaurs. When measurement of naturally occurring phenomena provides enough evidence that supports a theory and none that refutes it, causal statements are made and accepted, regardless of the scientist's inability to control the phenomena under investigation.

There are also many problems in psychology which cannot be studied experimentally. Every researcher faces ethical issues; for certain questions, ethical considerations may rule out the use of experimentation altogether. We do not feel free to assign subjects at random to conditions that might damage them physically or psychologically. For moral reasons, we are unwilling to tell someone that he or she is about to die, to remove people's frontal lobes, or to study the causes of psychosis by driving people crazy. Instead, we must find people who are already suffering in these situations and look to see if they are also characterized by other variables that we believe to be causes or consequences. For example, cigarette manufacturers frequently claim that there is no direct evidence showing that cigarette smoking *causes* lung cancer or heart disease in humans. They are right. This is because researchers are loath to randomly assign humans to "smoking" and "nonsmoking" conditions in order to demonstrate to skeptics that smoking cigarettes endangers life. Instead, they simply look to see if people who smoke are more frequently diagnosed with lung cancer or heart disease than people who do not smoke. Actually, in the case of research on smoking, researchers have done much more than "simply look." The simple observation was made decades ago, and hundreds of studies were conducted in attempts to separate the effects of smoking from the effects of diet, stress, personality type, and a host of other variables. The Surgeon General finally found the evidence convincing enough to declare that smoking *causes* the diseases that smokers get more often than nonsmokers. Without the experimental method, the demonstration of causal relationships generally requires a great deal of patience and effort and is often costly.

Studies in which the investigator is looking for a relationship between two variables which can be measured—but which cannot be controlled—are called **correlational studies.** When one variable causes another, the two will inevitably be correlated. However, the reverse is not true; discovering that two variables are correlated can never (without other evidence) provide unequivocal proof of causality. We might hypothesize, for example, that children who watch hours of violent television shows every day are more violent and aggressive than children who don't. Since we cannot assign some children at random to homes that encourage them to watch violent television shows and others to homes that prohibit them, we may decide that a correlational study is the most suitable method for studying this hypothesis. In such a study, we might observe children from 100 households, measuring the amount of violent television they watch. We might also measure the children's aggression: the frequency with which they fight with their siblings, break toys, throw tantrums, and so on. Suppose that our hypothesis was confirmed and that on the whole, the children who watched the most violence on television were also the most aggressive (a high positive correlation). Would we be able to conclude from these data that the children's aggression was *caused* by the violent television shows? No. There are all sorts of other things that could have led to the same high positive correlation.

First, it is possible that the causal relationship is exactly the opposite of the one we have hypothesized. Perhaps children who are more aggressive *to begin with* prefer watching television characters punch, shoot, stab, or otherwise harm one another more than do children who are less aggressive. An even

more common situation is that of a third-variable correlation; the two observed variables—*X* and *Y*—are correlated because both of them are highly correlated with (and maybe caused by) a third variable—*Z*—which we don't know about and haven't measured. Returning to our aggressive children, a third variable would exist if some other factor encouraged them both to view violent television shows and to behave aggressively. For example, it is possible that in homes where parents are frequently absent children have the opportunity to watch a lot of violent television shows and to engage in aggressive, destructive acts with no restraints. That is, the observed positive correlation may occur because parents are not at home to impose limits on television watching and aggressive behaviors. Or perhaps the relationship exists because parents who are themselves aggressive allow their children to watch violent fare on television, use corporal punishment, and condone aggressive behavior in their children. Plausible explanations for the observed positive correlation are limited only by our imaginations.

Thus, in a correlational study we can never be sure what the independent and dependent variables are. A correlational study can neither guarantee that the causal variables have been isolated nor provide the experimenter with control over the phenomena being studied. A correlational study does, however, allow the researcher to find out whether the phenomenon is *predictable* from knowledge about some other variable. Converging evidence from a large number and variety of correlational studies, all of which provide support for the same general theory, can ultimately satisfy scientists about causal relationships, as in the cases of eclipses and dinosaurs.

Basically, a correlational study consists of two sets of measurements, one of each member of a pair of variables which the scientist believes to be related: height and weight, extent of brain damage and extent of language loss, college board scores and grades in college, self-confidence and attractiveness, and so on. The degree of relationship between the two variables is tested mathematically and expressed as a **correlation coefficient** (symbolized as *r*). The value of *r* ranges from +1.00 to −1.00. A perfect **positive correlation,** +1.00, indicates that as one variable increases a constant amount, the other does too; by knowing how much of one variable is present, one can predict the exact value of the other. An *r* of −1.00 (a perfect **negative correlation**) also allows one to make exact predictions, but in this case the higher the value of one variable, the lower the value of the other (e.g., the *higher* the horsepower of a car, the *lower* the amount of time necessary to reach 60 miles per hour). An *r* of 0 indicates no relationship; knowing the value of one variable does not tell us anything about the value of the other (e.g., horsepower and color of the car). As *r* moves from 0 to +1.00 through the range of positive correlations or from 0 to −1.00 through the range of negative correlations, our ability to predict improves; the variables are more strongly related.

Very seldom do correlations among variables of interest to social psychologists approach +1.00 or −1.00. To cite but two examples, the correlation between general attitudes toward birth control and use of the birth control pill is only +.08 (Davidson & Jaccard, 1979) and the correlation between environmental attitudes and recycling behaviors is +.39 (Heberlein & Black, 1976). Although the magnitude of these correlation coefficients is not impressive in an absolute sense, it is worth noting that when the implications of even a "small" relationship are important, it is often worthwhile to pursue research. For example, the correlation between watching television violence and aggressive behavior ranges from +.10 to +.20 across a wide range of studies (Freedman, 1988): a relatively "weak" relationship. However, few people would argue that violence and aggression are not serious social problems. Thus, it is of both theoretical and practical import to determine if viewing television violence actually causes aggression. Moreover, if indeed it does, it is also important to determine how to explain that causal relationship and to arrive at ways to offset the deleterious effects of televised violence on aggressive, socially destructive behavior. Complex social variables such as aggression have many causes. It is completely unreasonable and simplistic to expect any single predictor variable to have a very high correlation with aggression. Even a history of actual aggressive behavior is an imperfect predictor. It is worthwhile, though difficult, to pursue even weak correlations if they keep turning up to find out whether a variable such as viewing televised violence is part of the reason some people are violent. The problem, of course, is to distinguish between weak-but-reliable correlations and imaginary ones.

Quasi-Experimental Designs

Of course, in the world outside the laboratory, there *are* phenomena that are easier to control than the movement of the planets, dinosaurs, or even what children watch on television. Social psychologists frequently design and implement interventions in government, industry, and educational settings and measure the effect of those interventions. However, despite control over the treatment or intervention, researchers are sometimes unable to assign individuals randomly to experimental and control conditions; they may be required to utilize only one group or to use intact, preexisting groups.

Studies that have independent variables and dependent variables but do not use randomization to assign subjects to groups are called **quasi-experiments** (Cook & Campbell, 1979). Because random assignment is not used, the groups compared are likely to differ in many ways aside from the treatment. In the absence of the "great equalizer," the groups are **nonequivalent,** and the researcher must determine which differences, if any, between these nonequivalent groups are due to the effects of the treatment and which are due to other differences between the groups. Without random assignment, one cannot say with certainty that the independent variable is the sole cause of some outcome, so the investigator has to "play detective" and logically and/or statistically eliminate as many rival causes as possible.

There are many varieties of quasi-experiments. Sometimes researchers use only one group to measure the effects of their intervention. Measures might be

ken at some time before an intervention and again at one or more points after
e intervention. Participants' behaviors or characteristics *before* the treat-
ent can then be compared to their behaviors or characteristics *after* the treat-
ent. For example, medical researchers may measure the physical and psy-
hological functioning of a group of patients *before* the administration of a new
rug and at various times *after* the administration of the drug. This strategy
llows for a comparison of the patients' functioning before and after the drug is
rescribed. Should symptoms subside, the researchers can be *somewhat* con-
ident that the drug was the cause of improvement.

Or, unable to create groups by random assignment, a researcher may use
wo groups that already exist, giving the treatment to one but not to the other.
These two groups can then be compared on dependent variables designed to
neasure the effects of the treatment. For example, researchers may be inter-
ested in improving employee morale in a corporation. Two different depart-
ments within the organization can be used to assess the effects of an inter-
vention designed to improve employee morale. Before the treatment is
implemented in one department, the morale of employees in both departments
is measured (e.g., attitudes toward the organization, ratings of job satisfaction,
absentee rates). Thereafter, employees in one department receive the
"morale-boosting" treatment, and employees in the comparison group carry
on as usual. The effectiveness of the treatment is then assessed by administer-
ing the same dependent measures to employees in both departments after
those in the experimental group receive the treatment. If employees exposed
to the treatment show improved morale compared to employees in the com-
parison group, results are *suggestive* of the efficacy of the intervention. We
hasten to add, however, that such a conclusion is still open to debate, because
we can't be sure that it was our intervention *alone* that caused improvement in
employee morale.

Whatever form they take, quasi-experiments differ from true experiments in
that subjects are not randomly assigned to treatment or control conditions. In
Chapter 5, we will provide a more detailed description of situations in which
quasi-experiments are either necessary or preferred, examples of various
quasi-experimental techniques, and obstacles to valid causal inferences asso-
ciated with them. For now, let us return to the experiment.

THE ADVANTAGES OF EXPERIMENTATION

The implication that the experiment is to be preferred over other techniques
has provided an undertone to much of our discussion so far. It is now time to
examine this implication explicitly and to ask why one should bother to at-
tempt an experiment in the first place. Certainly, there are disadvantages to
the social psychology experiment. It is often difficult to design. Hours of crit-
ical and creative thought go into selecting the appropriate empirical realiza-
tions of the experimenter's concepts. Alternative explanations must be elimi-
nated, and stimulus materials and dependent measures must be carefully

selected or developed. It is also likely to be laboriously time-consuming. Per-
mission to proceed must first be secured from human subjects committees;
subjects must be recruited; numerous "dress rehearsals" must be conducted
with pretest subjects to ensure that the instructions are understandable and
that the stimuli and events are interpreted appropriately. Once the experiment
is finally up and running, it is not unusual for the experimenter and one or
more assistants or confederates to spend an hour or more with each subject.
The experimenter frequently has to make elaborate preparations to set the
stage, to motivate the subject, and (in certain kinds of experiments) to deceive
the subject. After expending all this time and effort, the investigator may ob-
tain only a single datum: perhaps something as simple as a yes or no answer to
a single question. Once the data have been collected—questionnaire re-
sponses, subject ratings of stimulus materials, overt behaviors, and so on—
they must be scored, coded, and prepared for statistical analysis.

Furthermore—and this is perhaps the most common criticism—the experi-
ment is usually far removed from the real-life phenomena in which the exper-
imenter is supposedly interested. To the layperson it may seem ludicrous for
psychologists interested in the formation and change of basic attitudes and im-
portant values to study children picking out toys or eating spinach, or college
students guessing the lengths of lines, deciding how much they like a group on
the basis of an "interaction" consisting entirely of written messages, or using
7-point scales to rate a person who appears on a videotape.

Those of us who conduct experiments firmly believe that these disadvan-
tages are outweighed by the benefits of experimentation. In attempting to ex-
plain the reasons for this belief, we will examine one laboratory experiment in
some detail and compare it with other approaches that might have been used to
answer the same question. For illustrative purposes we have chosen the clas-
sic experiment by Elliot Aronson and Judson Mills (1959) which demonstrates
not only the advantages of the experimental approach but some of the pitfalls
as well. Aronson and Mills set out to test the hypotheses that a person who
undergoes a severe initiation in order to be admitted to a group will find the
group more attractive than if little or no initiation were required. To test this
hypothesis, they conducted the following experiment.

Sixty-three college women were recruited as volunteers to participate in a
series of group discussions on the psychology of sex. This format was a ruse,
created in order to provide a setting in which subjects could be made to go
through either mild or severe initiations in order to gain membership in a
group.

Each subject was tested individually. When a subject arrived at the labora-
tory, ostensibly to meet with her group, the experimenter explained that he
was interested in studying the "dynamics of the group discussion process"
and that, accordingly, he had arranged these discussion groups for the purpose
of investigating these dynamics, which included such phenomena as the flow
of communications and who speaks to whom. He explained that he had chosen
"The Psychology of Sex" as the discussion topic in order to attract a large

number of volunteers and that this had proved to be a successful device, since many college students were interested in this subject. There was, however, one major drawback: Many of the volunteers were embarrassed and found it more difficult to participate in the discussion than they might have if the topic had been a more neutral one. He explained that his study would be impaired if any group member failed to participate freely. He then asked the subject if she felt she could discuss this topic without difficulty. The subjects invariably replied in the affirmative.

These instructions were used to set the stage for the initiation. The subjects were randomly assigned to one of three experimental conditions: a severe-initiation condition, a mild-initiation condition, or a no-initiation condition. The no-initiation and the mild-initiation conditions constituted the control groups. As soon as the subjects in the no-initiation group told the experimenter that they had no qualms about discussing sex, they were told that they could join a discussion group. It was not that easy for the subjects in the other two conditions, however. The experimenter told these subjects that he had to be absolutely certain that they could discuss sex frankly before he could admit them to a group. Accordingly, he said there was a special test which he was using as a screening device to eliminate those women who would be unable to engage in such a discussion without undue embarrassment. In the severe-initiation condition, this embarrassment test consisted of having each subject read aloud (to the male experimenter) a list of twelve obscene words and two vivid descriptions of sexual activity from contemporary novels. In the mild-initiation condition, the women were merely required to read aloud a list of relatively inoffensive words related to sex. This elaborate procedure constituted the empirical realization of the independent variable.[2]

Each of the subjects was then allowed to "sit in" on a group discussion, being carried on, she was told, by the members of the group she had just joined. This group was described as one that had been meeting for several weeks; the subject was told that she would be replacing a group member who had to leave because of a scheduling conflict.

To provide all subjects with an identical stimulus, the experimenter had them listen to the same tape-recorded group discussion. At the same time, the investigators felt that it would be more involving for the subjects if they were made to believe that this was a live, ongoing group discussion. In order to accomplish this while justifying the lack of visual contact necessitated by the tape recording, the experimenter explained that since people found that they could talk more freely if they were not being looked at, each subject sat in a separate cubicle, talking through a microphone and listening through headphones. Since this explanation was consistent with the other aspects of the cover story, all the subjects found it convincing.

[2] Remember that this experiment was carried out in the mid 1950s when female undergraduates were unaccustomed to seeing (much less verbalizing) explicit obscene words and descriptions of sexual activity. As in most social psychological research, the independent variable must be understood in the cultural context of the time and place, a point discussed more fully later.

It was important to discourage the subject from trying to "talk back" to the tape, since if she did so she would soon realize that no one was responding to her comments. In order to accomplish this, the experimenter explained that it would be better if she didn't try to participate in the first meeting, since the other group members had done some reading for the meeting, and therefore the subject would not be able to participate on an equal footing. He then disconnected her microphone.

At the close of the discussion, the experimenter returned and explained that after each session, all members are asked to rate the worth of that particular discussion and the performance of the other participants. He then presented her with a list of rating scales. This was the measure of the dependent variable. The results confirmed the hypothesis. The women in the severe-initiation condition found the group much more attractive than did the women in the mild-initiation and the no-initiation conditions.

This experiment certainly has some of the disadvantages mentioned earlier. Most striking is the fact that the experimenters constructed an elaborate scenario bearing little relation to the real-life situations in which they were interested. The "group" which the subject found more or less attractive was, in fact, nothing more than a few prerecorded voices coming in over a set of earphones. The subject was not allowed to see her fellow group members or to talk with them. This is obviously a far cry from group interaction as it occurs outside the laboratory. Moreover, the use of deception raises both ethical problems and more pragmatic questions, such as whether the deception was successful.

The authors' hypothesis could have been investigated more directly and perhaps more easily by employing nonexperimental methods. For example, one might try to study it cross-culturally, rating the severity of the initiation rites into manhood in different cultures and correlating these with some index of the extent to which adult males like their group. A still more direct and perhaps simpler method would be to do a correlational study of existing fraternities. One might first observe whether the fraternities required initiations for membership. If initiations were required, one might rate them for severity. At a later time, one could return and interview the members of the various fraternities to find out how much they liked one another and their particular fraternity. If it turned out that the men in the fraternities requiring severe initiations liked their group better than did those in other fraternities, this would seem to provide far greater support for the hypothesis, since the evidence would have been gathered in a natural setting.

Unquestionably, this procedure has certain advantages. First, it is simpler than a laboratory experiment. It is unnecessary to recruit volunteers, tape-record a discussion, or go through an elaborate rigmarole designed to deceive college women. In addition, it is the real thing. Rather than a series of separate individuals, each listening alone to a tape recording of disembodied voices for

a short time, the fraternity situation involves real people living together in real groups over a relatively long time, developing strong positive or negative feelings toward one another. Moreover, there would be little question that the initiation we label as severe would in fact be a severe initiation. In the most extreme instances, the initiation would most certainly be far more severe than anything we could attempt in the laboratory.

However, there are some problems with this approach. First, the stimulus objects, that is, the fraternities themselves, vary a good deal in their inherent attractiveness. The severity of initiation, although hypothesized to be a cause of attractiveness, is certainly not the only cause. Obviously, groups have many characteristics which people find more or less attractive. Some groups are attractive because their members are friendly, perceptive, intelligent, athletic, "nice," generous, handsome, witty, and so on. Others are unattractive because members are dull, stupid, too loud, too quiet, too outgoing, too inhibited, and so on. In such a complex stimulus situation, the severity of initiation, although important, might be only one drop in a large bucket. Thus, because there is great variation among the fraternities in these and other attributes, it might be very difficult to demonstrate differences between severe-initiation fraternities and mild-initiation fraternities, even if initiation were the most important single determinant of attractiveness. In terms of our discussion earlier in this chapter, all these attributes of the fraternities are extraneous sources of random error. Since these sources of random error will undoubtedly affect the dependent variable—a liking for the fraternity—they can act to obscure or distort the effect of the independent variable, severity of initiation.

One of the great advantages of the experiment is that in the laboratory, the experimenter can often exert a great deal of control over such extraneous variables and thus ensure that the stimuli in the experimental conditions are similar. Thus, in the initiation experiment described above, the group whose attractiveness was to be judged was identical for all subjects. By holding constant all aspects of the group, Aronson and Mills succeeded in eliminating all the extraneous factors which may cause one group to be more attractive than another. In other words, they succeeded in minimizing random error, markedly increasing the odds that they would be able to successfully detect the effects of their independent variable, severity of initiation. Thus, although some degree of realism was sacrificed, one of the great gains was the achievement of considerable control over extraneous variations in all characteristics of the group to be rated. By reducing noise, they were better able to detect the signal.

This control, although highly desirable, is not in itself the major advantage of an experiment. There is one advantage that is far more important—the random assignment of experimental units (subjects) to experimental conditions. Let us suppose that the extraneous variation mentioned above was not great enough to obscure the relationship between severity of initiation and attractiveness. In other words, suppose we conducted a study of existing fraternities and discovered that the members of severe-initiation fraternities did find one

another more attractive than did members of mild-initiation fraternities. If this occurred, we would have to consider the possibility of alternative explanations.

The inability of such a correlational study to specify causes and effects is a fundamental weakness. If the hypothesis is supported by the data, the experimenter usually wishes to assert as a conclusion to the study that the independent variable caused differences in the subjects' behavior—in the Aronson and Mills experiment, that severe initiations *cause* increased liking for the group into which one is initiated. In our fraternity example there are a variety of other possible, and indeed plausible, explanations which involve different causal sequences. The simplest explanation for these results might involve a relationship which is the reverse of the one we have proposed. Rather than severe initiations causing high attractiveness, it may be that high attractiveness causes severe initiations. The more attractive group may perceive that they are attractive and may attempt to maintain this pleasant situation. Perhaps out of a desire to prevent the group from being diluted, they may try to discourage applicants and make it difficult for people to get into the group by requiring a severe initiation. Or perhaps only highly desirable fraternities can afford to require severe initiations, because only they can be sure of getting enough applicants who are willing to put up with it. One could list many other reasons why the attractive groups might tend to have more severe initiations. The point is that any such reason points to an explanation for our data which involves a causal sequence that is the exact opposite of the one we hypothesized. Since this study necessitates the investigation of groups that were in existence before we arrived on the scene, there is no clear way to determine from these data which of the two hypotheses is correct.

This analysis should provide a clear understanding of what is meant by the statement that correlation does not prove causation. Whenever we observe that variable X (e.g., severity of initiation) is correlated with variable Y (e.g., attractiveness of a group) we cannot be sure whether X caused Y or whether Y caused X. In our laboratory experiment, of course, there is no ambiguity. We know what caused X—the experimenter. Consequently, in observing that Y is correlated with X, the experimenter can be certain that Y cannot have caused X; X must have caused Y.

But let's return for a moment to the study of fraternities. We might decide to circumvent the above problem by actually intervening in the initiation rites of two existing fraternities by conducting a quasi-experiment. We might first find two comparable fraternities which typically practice relatively severe initiations. Further, we might even be successful at persuading members of one of these two fraternities to reduce the severity of their typical initiation rites. The result would be two fraternities—one that continues to use *severe* rites of initiation and the other that now uses *mild* rites of initiation. Now that we have two (nonequivalent) groups, we can compare new members' ratings of attraction to their respective fraternities. If members who experienced severe initiations like their new fraternity more than those who experienced mild ini-

tiations, can we justifiably conclude that differences in attractiveness were due solely to our treatment? Not necessarily. First, our two fraternities no doubt differ on dimensions aside from our initiation intervention. Moreover, although our quasi-experiment has provided us with more control than a correlational study, we still were unable to *randomly assign* prospective fraternity members to one or the other fraternity, and therefore we cannot be absolutely certain that the members of fraternity A did not differ from the members of fraternity B in some systematic way that we didn't anticipate.

Let us elaborate upon that point: Recall from earlier in this chapter that in correlational studies, there is the possibility that an observed correlation between X and Y is simply produced by some third variable which affects both of them. The same problem holds for our quasi-experiment above: There may be other variables that are correlated with both rites of initiation and members' liking for their fraternity. Again, the true experiment circumvents this pitfall. In an experiment, it is extremely unlikely that an adventitious third variable is correlated with the two variables under consideration. The reason for this is apparent when we look at the defining characteristic of an experiment. In an experiment, the experimenter both has control over what treatment each subject receives and determines that treatment by assigning subjects to conditions at random. If the subject receives a treatment that has been determined truly at random, it is virtually impossible for a third variable to be associated with that treatment. Consequently, such a third variable cannot affect the dependent variable. In the real-life fraternity example, the demon which is of constant concern to the investigator is the possibility that the independent variable, severity of initiation, has not been randomly assigned to the various subjects. Insofar as some unknown third variable affects applications to a fraternity with severe-initiation practices, that third variable might also affect how attractive the group will be to its new members.

One plausible third-variable correlation in the real-life fraternity example involves differences in the amount of the subjects' initial motivation. For example, it is reasonable to assume that some people may simply want to be in a fraternity without caring much about *which* fraternity, whereas other people will be motivated to join a specific fraternity, perhaps because they have reason to feel that they will be happier with the members of that group. If a specific fraternity has a reputation for requiring severe initiations prior to admission, those people who have a strong desire to join that particular fraternity will be willing to go through the initiation in order to join. However, people who simply want to belong to a fraternity, with no particularly strong feelings about which one, will be much more likely to choose a fraternity that imposes little or no initiation. After all, if a man does not care which fraternity he joins, why would he bother to go through a severe initiation to get into a particular fraternity? Consequently, a fraternity that requires little or no initiation will initially attract many people who have no special desire to be in that specific fraternity as well as some people who do have a great desire to be in that specific fraternity. However, a fraternity that requires a severe initiation will *primarily* attract those people who have a strong desire to be in that specific fraternity, a desire strong enough to allow them to endure the initiation. Therefore, any relationship between attractiveness and severity of initiation may be strictly a function of a disproportionate number of highly motivated people joining the severe-initiation fraternity. This problem is averted by the experiment. Random assignment of subjects to conditions not only guarantees that no unknown variable is causing the severe initiations to be administered by only one particular kind of group but also prevents the possibility that systematic motivational differences or personality characteristics among the potential joiners will cause the observed relationship.

Finally, in most laboratory experiments, one can vary the independent variable in a systematic manner, thus allowing for the isolation and precise specification of the important differences. If one were to study fraternities with different initiations, the likelihood would be that the various initiations would differ both qualitatively and quantitatively on a large number of dimensions. Suppose that the fraternity requiring a severe initiation asked its pledges to perform many demeaning jobs, to wear funny clothes, to submit to severe physical punishment, to expose themselves to danger, and to eat insects. In order to be sure which aspect of this complex treatment was causing the increased attractiveness, it would be ideal to have another fraternity that asked pledges to do all these things *except* submit to severe physical punishment. Instead, the ideal second fraternity would ask pledges to submit to mild physical punishment. Such a fraternity probably does not exist, but if we were creating initiations experimentally, we could produce an appropriate fraternity identical in all respects to the other one, except severity of physical punishment.

In sum, the major advantage of an experimental enquiry is that it provides us with unequivocal evidence about causation. Second, it gives us better control over extraneous variables. Finally, it allows us to explore the dimensions and parameters of a complex variable.