For demonstrating cause-effect relationships, the controlled experiment possesses certain advantages unequaled by other research methods, but no one method is best in an absolute sense. One's choice of method should be dictated by both the nature of the research question and the phenomenon under investigation. Moreover, to the extent that results obtained using numerous and varied methods—experimental, quasi-experimental, correlational—all converge, our faith in the validity of our inferences about social behavior is enhanced.

WHEN NONEXPERIMENTAL METHODS ARE DESIRABLE

Despite the advantages of experiments in enabling researchers to make unequivocal causal inferences, other considerations sometimes dictate forgoing the controlled experiment in favor of alternatives. First, as we mentioned above, some variables cannot be experimentally varied. Even though they are not amenable to experimental investigation, these variables may be profoundly important in helping us to predict and understand human behavior and might interact with other variables in which the researcher is interested. For example, in a recent review of research, Alice Eagly and Valerie Steffen (1986) found evidence for a small but consistent relationship between gender and aggression—overall, men tend to be slightly more aggressive than women. However, the influence of gender depends on the context within which aggression occurs. To cite but one example, although men are somewhat more aggressive than women in general, the difference is more pronounced in the case of physical aggression than verbal aggression. Of course, because subject gender cannot be manipulated in the laboratory, one cannot justifiably say that gender *per se* causes some people to behave more aggressively than others. Nonetheless, social psychologists who have questions about variables that predict aggression or other social phenomena need not—and should not—abandon research because some of those variables cannot be experimentally manipulated. Instead, for *some* questions, these researchers must use alternatives to the controlled experiment.

Other variables are out of bounds to social psychological researchers for ethical reasons. An experimenter cannot ethically manipulate love or hate, alcoholism, psychosis, or any of a large number of other things that are believed to be extremely important in motivating human social behavior. However, the researcher may believe that some of these variables are related to problems so pressing that they require immediate attention and that it is important to gather relevant information, even though it will not be possible to use experimental techniques that permit causal inferences.

Some processes operate over a longer time than is typically available to a laboratory researcher. Examples are the process of psychotherapy, the development of language in children, and the dynamics of a group that has predicted the end of the world a year from today. Researchers may be interested in these natural processes for their own sake or as instances of the operation of theoretically interesting variables. In either case, they would be unlikely to have much confidence in the results of an experimental analog that lasted only an hour or two.

Sometimes social psychologists adopt alternatives in studies conducted in the real world to reduce the chances that their subjects will feel like "guinea pigs" and therefore behave differently than they would in their normal interactions. In laboratory studies, it is often possible to disguise the true purpose of the experiment from the subjects, but they almost always know that they are in an experiment. Some kinds of behavior—particularly unflattering behavior—though common enough in other situations, may be very hard to elicit in the laboratory. Subjects are motivated not to show signs of stupidity, pettiness, recalcitrance, or aggression when they know that the experimenter is observing them. In a natural setting, subjects may be quite unaware that they are being studied at all, and so it is possible to rule out the hypothesis that a certain behavior is laboratory-specific. Of course, these natural settings do not always allow the researcher to conduct a true experiment.

Finally, many researchers use alternative methods as a means of suggesting, clarifying, refining, or extending experimental research findings. Often a broad correlational study will suggest hypotheses that can be tested and variables that later can be manipulated experimentally. Thus, *before* an experiment is run, correlational research can generate many ideas that can be further explored in more controlled conditions. *After* an experiment has been run, the researcher may want to return to the field to test out some implication of the findings in a natural setting. At this stage, a carefully conducted study may indicate other variables that mitigate the effects of variables isolated in the lab and thereby provide some information—often of a cautionary nature—about the generality of the results of the experiment.

An excellent example of the alternation between lab and field as a fruitful research technique is provided by a series of classic studies of group cohesiveness conducted by Leon Festinger and his colleagues at M.I.T. In the laboratory, cohesiveness was varied by appropriate descriptions of the experimental groups as highly attractive (Back, 1951) or highly cohesive (Schachter, 1951). Outside the laboratory, the researchers studied an M.I.T. housing project (Festinger, Schachter, & Back, 1950). One part of the project, called Westgate, had existed for some time, and within each court there was a great deal of communication among members and an overall liking for the group as a whole; the researchers defined these courts as cohesive units. The other part of the project—Westgate West—was newer, and the members seemed to act more as individuals than as group members; thus, the Westgate West living units were defined as low in cohesiveness. The research group asked the same kinds of questions in both the laboratory and the housing project and developed their theories on the basis of both types of research. In both situations the cohesive groups imposed greater pressure toward uniformity and tended to shun members who failed to go along with the group norms.

Sometimes, nonexperimental information is collected *during* an experiment. Even in the laboratory, the richness and complexity of the events that take

place in social interaction are often so great that many potentially important variables cannot be controlled. But if they are identified and measured, their relationships with the main variables of interest can be studied, and these relationships may suggest new possibilities for further experimentation. If the experimental prediction is not confirmed, the nonexperimental data may help the investigator to figure out whether the independent variables were misinterpreted or weak, the measure was inadequate, some extraneous variable interfered, or the original conceptualization failed to take into consideration some important aspect of the phenomenon. Such an analysis, called an **internal analysis**, cannot prove a causal relationship—it is still a correlational analysis, even if it takes place in the context of an experiment—but it can be extremely useful as a source of information for guiding future experimentation.

Even when an experimental prediction is confirmed, an internal analysis can provide useful information. For example, it may increase the experimenter's confidence that the independent variable resembles its real-world counterpart. In an experiment designed to find out whether people with low self-esteem or high self-esteem were more likely to cheat, Elliot Aronson and David Mettee (1968) not only *manipulated* self-esteem, assigning subjects to high, medium, or low self-esteem conditions at random, but also gave subjects a test that *measured* their preexisting, or "real," levels of self-esteem. The test was administered before the subjects were given the false personality-test information that constituted the experimental manipulation of self-esteem, so that their test scores could not be affected by anything the experimenter told them. The correlation between the real self-esteem scores and the amount of cheating tended to support the authors' results obtained on the basis of the experimental manipulation of self-esteem; either way, the people with low self-esteem cheated more that the self-confident subjects did. The highest percentage of cheaters was found in the group of subjects who were given negative personality feedback (manipulated self-esteem) *and* who had unfavorable self-concepts to begin with (measured self-esteem). The lowest percentage was found in the opposite group: the subjects who were given positive personality feedback and who had favorable self-concepts to begin with. The finding that low baseline levels of self-esteem tended to influence subjects in the same direction as the negative personality feedback they were given—in both cases, the subjects were more likely to cheat—strengthened the experimenters' confidence that giving subjects unfavorable feedback from personality tests was in fact affecting their level of self-esteem.

A corollary function of an internal analysis performed within the context of a successful experiment is to provide data that may have a bearing on the plausibility of an alternative explanation for the main results. For example, one alternative explanation of the results in the Aronson and Mettee experiment was that receiving a low score on the personality test made the subjects angry and that they cheated on the test or order to get revenge, perhaps intending to invalidate the results of the experiment. According to this explanation, the increased level of cheating was due not to low self-esteem but to anger. The data from the measurement of chronic self-esteem levels do not support this alternative explanation, however; the fact that subjects with "chronic" low self-esteem behaved in the same way as subjects whose self-esteem had been lowered by unfavorable personality test feedback suggests that self-esteem was in fact the important factor differentiating cheaters from noncheaters.

So far in our discussion of alternative methods, we have discussed them in general terms—as if they are all pretty much alike. In one sense, they *are* similar: Unlike the controlled experiment, none of them can provide unequivocal answers to questions about causal relationships, because only in an experiment can we control and systematically vary the independent variable and assign subjects at random. However, as we noted in Chapter 1, alternative methods vary tremendously in terms of how closely they approximate a true experiment. Some correlational studies tell us *nothing* about causal relationships between variables. But there are some quasi-experimental designs that afford researchers considerable control over experimental conditions and differ from experiments only in that individual subjects are not assigned to experimental or control conditions in a truly random fashion. Thoughtful and thorough statistical analyses of data generated in these quasi-experiments can allow researchers to make *tentative* statements about cause and effect.

CORRELATIONAL STUDIES

In Chapter 1 we discussed a number of research methods—experiments, quasi-experiments, correlational studies—that might have been used to discover whether people who undergo a severe initiation to join a group will be more attracted to the group than people who undergo a mild initiation or no initiation at all. And we noted a fundamental limitation of a correlational study: the inadequacy of such an approach in establishing a causal connection between severity of initiation and liking for the group. That is, Aronson and Mills (1959) might have chosen to study existing fraternities, and rated initiations for severity. Later, they might have returned to interview new members of the different fraternities and had the members rate how attractive they found their respective fraternities. Such an approach would have yielded an index (a correlation coefficient) of the strength of the relationship between the two variables but would have failed miserably in determining the exact nature of that relationship. For example, severe initiations might cause new members to rate their fraternities as more attractive. However, fraternities that are highly attractive to begin with might have to resort to severe initiations to limit membership to only the most desirable of pledges; that is, the fraternities' attractiveness might cause severe initiations. This is an example of the direction of causality problem inherent in many kinds of correlational research. No less problematic is the possibility that some third variable—such as pledges' initial motivation to join *any* fraternity as opposed to *a certain* fraternity—might be responsible for the observed relationship between severity of initiation and liking for the group. Thus, the major disadvantage of the correlational study described above is that it does not allow researchers to make definitive statements about just what causes what.

This does not mean that all correlational studies are useless in enabling researchers to make tentative causal inferences. Indeed, some correlational designs offer a *partial* solution to the directionality problem we described above. One such design, labeled by Donald Campbell and Julian Stanley (1966) the **cross-lagged panel technique**, provides just such a partial solution. Predicated on the notion that "causes" precede "effects," the cross-lagged panel design involves collecting correlational data on at least two separate occasions. That is, two variables assumed to be related are observed at time 1 and again at time 2.

For example, Monroe Lefkowitz and his colleagues (Lefkowitz, Eron, Walder, & Huesmann, 1972) were interested in the relationship between the amount of violence children watch on television and their tendency to behave aggressively. Rather than conduct a laboratory experiment, they chose to study behaviors in the real world using a correlational investigation. Initially, these researchers measured the relationship between the violence contained in third-grade children's favorite television shows and their aggressiveness in the classroom. And they *did* find a relationship between expert ratings of violence in the children's preferred television shows and peer ratings of classroom aggression. Had Lefkowitz and his colleagues terminated their research at this point, it would not have been informative with regard to the *causal* relationship between watching television violence and behaving aggressively. As we noted in Chapter 1, it is possible that children who are aggressive to begin with are more likely than other children to prefer violent television programs; that is, initially high levels of aggression might cause children to prefer violent TV fare. Alternatively, a steady diet of TV violence might cause children to behave aggressively. Therein lies at least one source of ambiguity in a simple one-shot correlational study.

Aware of that problem, Lefkowitz and his colleagues did not stop there. Ten years later, they were able to contact about half of their initial subjects—quite a feat, given the passage of a decade. Once again, the researchers obtained ratings of the violence content of subjects' preferred television shows and peer ratings of their aggressiveness, and they were able to calculate the strength of the relationship between those two variables. Armed with data from two sets of observations separated by a 10-year lag, the researchers could calculate additional correlation coefficients: (1) between TV violence ratings at time 1 and time 2; (2) between aggression ratings at time 1 and time 2; (3) between TV violence ratings at time 1 and aggression ratings at time 2; and (4) between aggression ratings at time 1 and TV violence ratings at time 2. Figure 5-1 provides a summary of those correlations for male subjects.

The correlation coefficients of interest are those on the diagonals: (1) the correlation between TV violence ratings at time 1 and aggression ratings at time 2, and (2) the correlation between aggression ratings at time 1 and TV violence ratings at time 2. The usefulness of these two correlation coefficients is based on the assumption that what comes sooner *might* cause what comes later, but that the reverse is impossible: What comes later cannot possibly cause what came before it. Thus, if TV violence in third-graders is correlated with aggression 10 years later, one might assume a causal relationship between those two variables. Similarly, if aggression in third grade is correlated with TV violence preferences 10 years later, one might assume a causal relationship between *those* two variables. To determine which of the two assumptions is accurate (watching violent TV in third grade leads to aggression 10 years later, or being aggressive in third grade leads to watching violent TV programs 10 years later), the researchers pit these two correlations against one another. Lefkowitz and his colleagues found that the correlation between TV violence ratings at time 1 and aggression ratings 10 years later ($r = .31$) was significantly higher than the correlation between aggression ratings at time 1 and TV violence ratings 10 years later ($r = .01$). Thus, they felt fairly safe in assuming that watching television violence causes aggression rather than the other way around.[1]

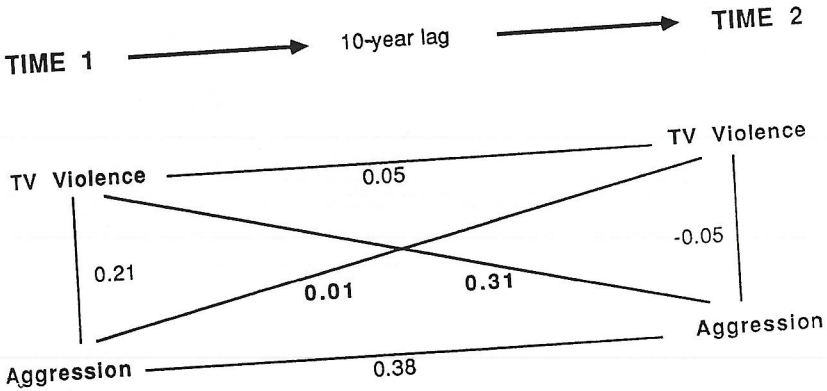

FIGURE 5-1
Correlations between television violence and aggressive behavior in boys. [From J. M. Neale and R. M. Liebert (1986). *Science and behavior: An introduction to methods of research* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall. (*Used by permission.*)]

[1] As we've noted, the cross-lagged panel design is only a *partial* solution for problems inherent in making causal inferences based on correlational data. First, it does not rule out the possibility that some third variable is responsible for an observed correlation. Second, before they can make even tentative assumptions about which of two correlated variables causes the other, researchers must be confident that they have obtained reliable measures, and that they have chosen an appropriate time interval between observations. Finally, a number of researchers have questioned the use of correlation coefficients as appropriate indices of association and have recommended more sophisticated statistical techniques (for example, see Neale & Liebert, 1986; Rogosa, 1980). Nonetheless, most researchers are in agreement about the validity of the underlying logic: Indices of association can be used to make tentative assumptions about causal links when the relationship between two variables is measured at more than one point in time. Furthermore, laboratory and field experiments, and subsequent correlational studies employing more sophisticated statistical techniques (e.g., Eron & Huesmann, 1980; Eron & Huesmann, 1986; Huesmann et al., 1984) have supported the initial conclusion of Lefkowitz and his colleagues.

Sometimes researchers are interested in being able to make at least tentative causal inferences based on correlational data. When that's the case, the cross-lagged panel technique is preferred over correlational studies in which observations are made at only one point in time. But not *all* research is aimed at establishing a causal relationship between two variables, and even the "simplest" of correlational studies can be valuable in formulating and evaluating theory. Theories specify certain relationships between phenomena. *One* way to evaluate the adequacy of a theory is to observe whether predicted relationships between these phenomena do indeed exist, by carefully examining variables of interest and determining if they are associated in ways predicted by the theory. Especially in the early stages of theory development and refinement, it is the task of researchers to discover whether there is a relationship between variable X and variable Y, period. Is there a relationship between birth order and achievement motivation? Is there a relationship between self-esteem and liking for others? Is there a relationship between frustration and aggression? Correlational studies are well suited to answering these kinds of questions. Once it has been established that theoretical predictions about the relationships between variables are accurate, other methods—such as the controlled experiment—can be used to test more specific causal assumptions. In essence, correlational studies are good first, but tentative, steps in evaluating theory.

Finally, there is one important thing correlational data can tell us about causality, and that is that *no* simple causal relationship exists. Although correlational data cannot prove causality, they can disprove it. If none of their correlations between television violence and aggression had differed significantly from zero, Lefkowitz and his colleagues would have had good reason to doubt that violent television has much of an effect on aggression. Sometimes complex combinations of other variables can mask a real correlation between the two the investigator cares about—perhaps even masking a true causal relationship—but even in these cases, the experimenter has to recognize that any causal relationship is more complicated than expected.

PSEUDOEXPERIMENTAL (PREEXPERIMENTAL) DESIGNS

As we have noted, nonexperimental designs vary in terms of how closely they approximate a controlled experiment.[2] Purely correlational studies, whether based on one observation or a series of observations, differ from experiments in a number of ways. One important way is that researchers who conduct correlational studies do not administer a treatment. Whatever the assumed "independent variable" (e.g., the amount of violence children watch on television), it is beyond their control. The researchers do not create the variables of interest but rather find a setting in which they occur naturally.

[2] The following discussion of various types of pseudoexperimental design is based on Donald Campbell and Julian Stanley's (1966) monograph, *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand-McNally.

In this section, we will discuss a number of designs that are closer approximations of the controlled experiment. These methods *do* involve an independent variable administered by the researcher. After subjects are exposed to this variable, its effects are assessed by measuring some dependent variable. However, these designs differ from experiments in that they do not contain appropriate standards of comparison (control groups or control observations) against which to measure the effects of the independent variable. Because of the lack of control inherent in these designs, numerous variables—in addition to the independent variable—might have produced an observed effect. Therefore, like correlational designs, these pseudoexperimental designs do not permit completely valid causal inferences about the effect of the independent variable on the dependent variable.

The One-Shot Case Study (Posttest-Only Design)

Suppose that you are interested in the effects of alcohol. Probably the simplest thing to do would be to invite a group of people in, give them all the dry white wine they wanted, and observe their actions. You might write down that they seemed very relaxed and contented, that they touched one another a great deal, and that they said they were feeling more relaxed and less inhibited than usual. What would you learn from your study? Much less than you might think.

We present this kind of study—in which one group is observed during or after an event that interests the researchers—merely as a point of departure. The most fundamental weakness of such a study is that it provides no basis for comparison, and comparison is essential to science. Without a basis for comparison, we have no information about similarities and differences. For example, suppose that a man goes from the city to visit his friend in the country and that while he is out walking, he sees an owl. He has seen pictures of owls before, but he has never seen a live one. When he gets back to the house, he tells his friend, who asks, "Was it a big one?" This question is unanswerable. Since the man has seen only one owl, he has no idea what's big for an owl; that is, he has no basis for comparison. He might say, "It was bigger than a pigeon," and this might enable the friend—who knows the size range of owls in the vicinity—to decide whether it was a "big" one. Similarly, if the friend from the country were to visit the city, she would be unable to say whether the subway made good time on her first ride.

One might object that this is a far-fetched example, since in the alcohol study the researcher was measuring how happy people were, how much they touched one another, and so on. It is not as though this were the researcher's first experience with happiness, or physical contact, or inhibitions; he or she has a sort of feeling for how groups of people usually behave when they come over—a whole storehouse of thoughts, sensations, and memories about past

parties. But a "sort of feeling" is really a very vague and nebulous basis for comparison, subject to all kinds of distortion from forgetfulness, personal bias, expectations about the results of the experiment, and so forth. A researcher who wants to compare two sets of measurements should make both sets of observations in the same way, with the same care. Even the most thoughtful conjecture about what the data would look like without the independent variable is no substitute for a standard of comparison.

The alcohol example is an obvious one, so transparently inadequate that you may well be grumbling that there is no point in wasting so much exposition on a design so stupid that nobody would ever think of using it to conduct research. Campbell and Stanley (1966, p. 6) call this type of design—in which a single group is observed on a single occasion—a **one-shot case study**, and surprisingly enough, it is not uncommon. One-shot case studies are not always easy to detect, because they are sometimes used when a vast amount of complex information about some person or group is collected and tabulated and analyzed and compiled into a presentation so erudite and detailed that the reader may lose sight of the fact that no other person or group was studied. An entire team of researchers, interested in how parental divorce affects child development, might study a sample of 500 or even 5000 children whose parents were divorced, measuring school performance, general psychological adjustment, number of friends, incidence of delinquent acts, bed-wetting, and a hundred other variables, and interviewing each child and each parent in depth. The task of collecting, tabulating, and organizing the data is immense, and the reader of the resulting 600-page book may find it difficult to remember that all this care and precision is misplaced because no comparison has been made. Conducting such a study is analogous to building a castle on quicksand. It is impossible to tell whether the children are different from the average for their social group on any of the variables. It is impossible to tell where they stood on all the measures before the divorce; maybe nothing has changed.

Is there ever any point to using this technique? Certainly, if nothing else is possible. Occasionally, one may happen on a once-in-a-lifetime situation, such as the Chernobyl disaster or the loss of the space shuttle *Challenger*, when there is no choice but to study it then and there as it occurs or to let it slip away into oblivion. When the choice is between letting an interesting phenomenon go by and studying it, even though the study may be an isolated analysis, it would be foolish to pass up the opportunity because no better methods are available. Our criticisms of various designs are meant not to discourage research but to *improve* it. In the hypothetical study of the effects of divorce, there was no apparent reason that the study could not have used a sample of 250 or 2500 children of divorced parents and an equal sample of children whose parents were not divorced, and this would have been a far better study. Sometimes, however, one is exposed to a special case: a patient with a well-differentiated case of multiple personality, a group of people who firmly believe that the world will end on a particular date in the near future, a cultural group (such as the kibbutz) that has drastically curtailed the role of the nuclear family. Anthropology is based on the in-depth study of individual cultures. A special case that does not conform to common behavior patterns in our society can provide valuable information about the range of human behavior.

Given the constraints of a one-shot case study, there are more or less adequate ways of carrying it out. If one is interviewing a person or a group of people after a disaster, one can ask them about their behavior in other situations; although this is not a good basis for comparison, it is better than nothing. In addition, if the researcher has a choice between situations that offer differing amounts of control over the variable whose effects are being studied, the researcher should generally choose the situation that can be more closely controlled. Thus, in the alcohol study, the fact that you, the investigator, handed out the wine makes this study somewhat superior to a correlational study in which you simply observed people who were already drinking when you walked in, since you had *some* control over the occurrence of the treatment. But if you have this much control, you can usually conduct an even better study.

One-Group Pretest-Posttest Design

To return to the alcohol study, suppose that you invited the group over at 8 P.M. and simply observed them from 8 until 10 (the *pretest* observations), handed round glasses of wine, waited an hour to be sure it had begun to take effect, and then conducted a second set of observations from 11 P.M. until 1 A.M. (the *posttest*). Now you have something to compare: the pre–10 P.M. observations with the post–11 P.M. observations. If you find that people are happier, more relaxed, and so on after drinking the wine, you know one thing that you couldn't know in the first study: It wasn't just that those people were in that state all along, even when they first walked in the door.

But this technique still leaves a good deal to be desired. First, you can't say that the increase in happiness and relaxation would not have occurred anyway, even if you hadn't served drinks. There are other things that could have caused the increase. Any number of things might have caused your subjects to change their behaviors between 10 P.M. and 1 A.M., and these other possibilities threaten your ability to make a valid causal inference about the effects of alcohol consumption on social behavior. Campbell and Stanley (1966) refer to these variables that are beyond the control of the researcher as **threats to internal validity**. They prevent the researcher from concluding that the independent variable was the sole cause of the outcome. Although this **one-group pretest-posttest design** (Campbell & Stanley, 1966, p. 7) is an improvement over the one-shot case study, a number of these threats to internal validity still exist:

1 Perhaps something else happened between the pretest and the posttest that caused changes in behavior. Perhaps just sitting around and talking with the group for a while would have resulted in greater relaxation and a lowering of inhibitions; conversation can most certainly loosen people up. Or perhaps someone told a joke, and the whole group cheered up and relaxed. These un-

controlled events that occur between the pretest and posttest might well be the cause of a change in behavior revealed in the posttest. These threats to a valid causal inference are called **history** effects and are a source of rival explanations for the apparent effects of the independent variable.

2 Or perhaps the guests acted relaxed because it was late and they were getting tired. **Maturation** refers to changes that unfold within subjects as a function of the passage of time between the pretest and the posttest, independent of any external events. Maturation can refer to any number of internal biological or psychological processes—growing older, growing tired, growing bored, growing hungry, or whatever—that provide a rival explanation for the effects of the treatment.

3 Perhaps your subjects saw you writing things down and didn't want to look uptight; or perhaps they guessed that you wanted to see if they acted inebriated after drinking the wine. When people are the objects of investigation, the very fact of being observed—and not the treatment—can be the cause of a change in their behavior between the pretest and the posttest. This threat to internal validity is called **testing.**

4 You observed and recorded people's behaviors over the course of a five-hour period. Perhaps you yourself were tired and relaxed from 11 P.M. to 1 A.M., knowing that the study would soon be over, and you wrote down your observations more casually or failed to notice small indications of tension in the room. Perhaps you'd indulged in a little of the independent variable yourself. Measuring instruments themselves—hardware or human observers—can change between pretest and posttest. Mechanical measuring instruments can malfunction or deteriorate. Human observers can become more experienced or more fatigued between the pretest and the posttest; their observations may become more accurate and reliable or more haphazard with the passage of time. Thus, differences between subjects' pretest scores and posttest scores might be due to changes in the **instrumentation** itself rather than the effects of the treatment.

5 Perhaps after drinking the alcohol, a number of subjects left the room unnoticed before you began your observations at 11 P.M. Of course, the subjects who left the room may not have differed from those who remained, but you can't be sure. Perhaps unlike those whom you observed growing more relaxed and convivial from 11 P.M. until 1 A.M., these subjects grew anxious or depressed after consuming alcohol and left to go recover; or perhaps the alcohol had no effect on them at all and they left because your party bored them. If the subjects who left the party differ systematically from those who remain, you can't be sure if differences between pretest and posttest scores are due to effects of treatment, or due to the fact that subjects who remain in the study are somehow different from the initial sample. This threat to internal validity is rather gruesomely known as **mortality**, because some subjects fall by the wayside.

All the above extraneous variables—specific interfering events, the passage of time, the presence of the tester, changes in the tester over time, or subject attrition—are confounded with the treatment and give us reservations about a too-confident belief in the effects of alcohol. It impossible to disentangle the effects of alcohol from the effects of anything else that might have changed between the first and second sets of measurements.

Again, it should be pointed out that this kind of design is not uncommon and that complicated and extensive observations often mask the simplicity and inadequacy of the basic design. The case study of a patient undergoing therapy is typically an example of this type of design. A schizophrenic's pretreatment behavior may be outlined in detail, with long historical accounts of the patient's childhood and of the parents' personalities. Then, the schizophrenic may be examined for years, during the course of a novel and complex therapeutic method (the treatment); the final write-up may contain fascinating interview transcripts, careful charts of day-to-day variations in a large number of the patient's behaviors, dramatic accounts of how the patient appeared to be sliding back into the abyss, and finally the excruciating step-by-step improvement that eventually culminated in the patient's triumphant release from the hospital. But all we know is that the patient got better. We don't know *why*; maybe the therapy had something to do with it, maybe not.

Once again, we should emphasize that when no more sophisticated design is feasible, it is often better to conduct research according to this simple one-group pretest-posttest design than to neglect an opportunity to collect interesting information. Leon Festinger, Harry Riecken, and Stanley Schachter (1956) used a design of this sort to study a group whose members had predicted the end of the world, and the general theory the researchers used to interpret their findings was subsequently supported by an extensive program of laboratory research. Taken in conjunction with the experimental studies, the study of the doomsday group provided important evidence of the applicability of the findings to real-life situations, as well as a great deal of suggestive data on the detailed dynamics of the process of dealing with an event that disconfirms one's most important expectation.

The Static-Group Comparison Design

In the first pseudoexperimental design, the one-shot case study, we pointed out that the only kind of comparison that the researcher could make was between the treatment group and remembered impressions of other groups or assumptions of what the group under study might have been like without the treatment. In the second design, the one-group pretest-posttest design, a comparison was afforded by measuring the same group before and after they had drunk alcohol. But this comparison was confounded with many changes that could have occurred as the night wore on: in the subjects (maturation), in the experimenter (instrumentation), and in the environment (history, testing, and mortality).

There is a second alternative in the choice of a standard of comparison: **a static-group comparison design.** Suppose that on the night you ran the original one-shot case study of the effects of alcohol, someone else in your building was having a party at which no alcohol was served. Knowing that your original

controlled events that occur between the pretest and posttest might well be the cause of a change in behavior revealed in the posttest. These threats to a valid causal inference are called **history** effects and are a source of rival explanations for the apparent effects of the independent variable.

2 Or perhaps the guests acted relaxed because it was late and they were getting tired. **Maturation** refers to changes that unfold within subjects as a function of the passage of time between the pretest and the posttest, independent of any external events. Maturation can refer to any number of internal biological or psychological processes—growing older, growing tired, growing bored, growing hungry, or whatever—that provide a rival explanation for the effects of the treatment.

3 Perhaps your subjects saw you writing things down and didn't want to look uptight; or perhaps they guessed that you wanted to see if they acted inebriated after drinking the wine. When people are the objects of investigation, the very fact of being observed—and not the treatment—can be the cause of a change in their behavior between the pretest and the posttest. This threat to internal validity is called **testing**.

4 You observed and recorded people's behaviors over the course of a five-hour period. Perhaps you yourself were tired and relaxed from 11 P.M. to 1 A.M., knowing that the study would soon be over, and you wrote down your observations more casually or failed to notice small indications of tension in the room. Perhaps you'd indulged in a little of the independent variable yourself. Measuring instruments themselves—hardware or human observers—can change between pretest and posttest. Mechanical measuring instruments can malfunction or deteriorate. Human observers can become more experienced or more fatigued between the pretest and the posttest; their observations may become more accurate and reliable or more haphazard with the passage of time. Thus, differences between subjects' pretest scores and posttest scores might be due to changes in the **instrumentation** itself rather than the effects of the treatment.

5 Perhaps after drinking the alcohol, a number of subjects left the room unnoticed before you began your observations at 11 P.M. Of course, the subjects who left the room may not have differed from those who remained, but you can't be sure. Perhaps unlike those whom you observed growing more relaxed and convivial from 11 P.M. until 1 A.M., these subjects grew anxious or depressed after consuming alcohol and left to go recover; or perhaps the alcohol had no effect on them at all and they left because your party bored them. If the subjects who left the party differ systematically from those who remain, you can't be sure if differences between pretest and posttest scores are due to effects of treatment, or due to the fact that subjects who remain in the study are somehow different from the initial sample. This threat to internal validity is rather gruesomely known as **mortality**, because some subjects fall by the wayside.

All the above extraneous variables—specific interfering events, the passage of time, the presence of the tester, changes in the tester over time, or subject attrition—are confounded with the treatment and give us reservations about a too-confident belief in the effects of alcohol. It impossible to disentangle the effects of alcohol from the effects of anything else that might have changed between the first and second sets of measurements.

Again, it should be pointed out that this kind of design is not uncommon and that complicated and extensive observations often mask the simplicity and inadequacy of the basic design. The case study of a patient undergoing therapy is typically an example of this type of design. A schizophrenic's pretreatment behavior may be outlined in detail, with long historical accounts of the patient's childhood and of the parents' personalities. Then, the schizophrenic may be examined for years, during the course of a novel and complex therapeutic method (the treatment); the final write-up may contain fascinating interview transcripts, careful charts of day-to-day variations in a large number of the patient's behaviors, dramatic accounts of how the patient appeared to be sliding back into the abyss, and finally the excruciating step-by-step improvement that eventually culminated in the patient's triumphant release from the hospital. But all we know is that the patient got better. We don't know *why*; maybe the therapy had something to do with it, maybe not.

Once again, we should emphasize that when no more sophisticated design is feasible, it is often better to conduct research according to this simple one-group pretest-posttest design than to neglect an opportunity to collect interesting information. Leon Festinger, Harry Riecken, and Stanley Schachter (1956) used a design of this sort to study a group whose members had predicted the end of the world, and the general theory the researchers used to interpret their findings was subsequently supported by an extensive program of laboratory research. Taken in conjunction with the experimental studies, the study of the doomsday group provided important evidence of the applicability of the findings to real-life situations, as well as a great deal of suggestive data on the detailed dynamics of the process of dealing with an event that disconfirms one's most important expectation.

The Static-Group Comparison Design

In the first pseudoexperimental design, the one-shot case study, we pointed out that the only kind of comparison that the researcher could make was between the treatment group and remembered impressions of other groups or assumptions of what the group under study might have been like without the treatment. In the second design, the one-group pretest-posttest design, a comparison was afforded by measuring the same group before and after they had drunk alcohol. But this comparison was confounded with many changes that could have occurred as the night wore on: in the subjects (maturation), in the experimenter (instrumentation), and in the environment (history, testing, and mortality).

There is a second alternative in the choice of a standard of comparison: a **static-group comparison design.** Suppose that on the night you ran the original one-shot case study of the effects of alcohol, someone else in your building was having a party at which no alcohol was served. Knowing that your original

study suffered from the lack of a comparison group, you might decide to run down the hall and make a few observations at the other party, where no one was drinking wine. The trouble with this plan is that you have no way of knowing that the alcohol had anything to do with differences you might find between your party and the other person's. Maybe your friends were livelier to begin with. That is, perhaps differences in the behavior of people at the two parties were not because alcohol was consumed at one party and not at the other, but because people at each of the two parties were different from one another at the outset. Campbell and Stanley (1966, p. 5) refer to this threat to internal validity as **selection**. Any time two or more comparison groups are used, and subjects are not assigned to those groups at random, selection poses a threat to valid causal inference.

Different kinds of guests might have attended the two different parties. Furthermore, the parties might have differed on other dimensions: Maybe the other person's furniture was less comfortable than yours, maybe the other person has a stomach ache and is sorry that people are coming over that night. There are many factors besides the alcohol that could create differences between the other person's party and yours. And selection may have *interacted* with these other factors to produce differences that would be erroneously attributed to the alcohol treatment. For example, maybe guests at your party react differently to being observed than guests at the other party because you're their friend, and they don't expect it of you (selection × testing interaction); maybe guests at your party got up later than guests at the other party, and don't tire as easily during the course of the evening (selection × maturation interaction); maybe people at your party are less likely than guests at the other party to leave before you make your observations (selection × mortality interaction). Because you can't know whether the guests at the two parties would have behaved identically if you hadn't distributed alcohol, you can't conclude that the differences you observe have anything to do with drinking. Thus, such a comparison group is not very useful, as it doesn't really "control" for very much.

Still, the fact that this is not the best type of control group should not prevent you from going down the hall and making observations at the other party. If it is the best control group you can find, it is better than nothing and should be used.

QUASI-EXPERIMENTAL DESIGNS

Some nonexperimental designs provide the researcher with more control than is afforded by the pseudoexperimental designs discussed above.[3] In these

[3] Once again, we have relied heavily on the excellent discussion provided by Campbell and Stanley (1966). Space limitations preclude a detailed discussion of *all* of the quasi-experimental designs discussed in their monograph. Interested readers are encouraged to consult their classic for further information.

quasi-experimental designs (Campbell & Stanley, 1966, p. 34), the researcher is able to control the administration of a treatment—for example, when it is administered, and to whom—and the collection of dependent variable data. The major difference between quasi-experiments and true experiments is that in the former, subjects are not assigned at random to experimental groups. As a result, quasi-experiments have one advantage over controlled experiments: They are more feasible research techniques when researchers move beyond the laboratory to applied or field settings in which random assignment of individuals to conditions is impossible. However, there is a trade-off, for the lack of random assignment robs the quasi-experimental researcher of the power to make unequivocal statements about cause and effect—statements that *can* be made when a true experiment is used. Even so, quasi-experiments are valuable research tools, for depending on the patterns of results obtained in these kinds of studies, researchers are able to make *relatively straightforward* causal interpretations.

There are two general kinds of quasi-experiments: In **interrupted time series designs**, the effect of the independent variable is assessed by comparing multiple observations of the same group of individuals, either before and after administration of one level of the treatment, or after the administration of different levels of a treatment. In essence, these quasi-experiments make use of a *within-subjects* design (see Chapter 4). In **nonequivalent control group designs**, the effect of the independent variable is assessed by comparing two or more groups of subjects—subjects who are not assigned at random to their respective groups. These quasi-experiments make use of a *between-subjects* design (see Chapter 4).

Time-Series Experiment (Simple Time Series Design)

Remember our one-group pretest-posttest design. In that pseudoexperiment, the experimenter would have preferred to make the observations from 8 to 10 P.M., then go back in time to 8 P.M., erasing everyone's memory of the intervening events, and start over. This, of course, is impossible, but a slight improvement in this direction can be made if the experimenter can observe the same group of people regularly—every day or once a week—at 8 P.M. No longer does the investigator have to conduct the pretest, introduce the treatment, and conduct the posttest all on the same night. Whenever the group comes over, the experimenter can observe people's behavior. After several nights of this, the experimenter will give the group alcohol and continue with the observations. The next few times they come, the researcher will continue to observe the group's behavior, again without handing out alcohol. The experimenter now has what Campbell and Stanley (1966, p. 37) call a **time-series experiment**.

Why not just invite the group twice, once giving them alcohol and once not? This procedure would eliminate such factors as the group members' fatigue over the course of the evening, but most of the problems of the pretest-

TABLE 5-1

Alcohol served	Day	Number of touches
No	Monday	4
No	Tuesday	3
No	Wednesday	5
Yes	Thursday	15
No	Friday	7
No	Saturday	4
No	Sunday	5

posttest design would still remain. For example, the group members might be more relaxed the second evening, because the situation is more familiar to them. We are more confident about drawing conclusions from a time-series study, because this type of study allows us to see how distinctive the alcohol night is, compared to a number of similar occasions without the alcohol.

A specific example will allow us to explain this advantage more clearly. Suppose that one of the measures is the number of times people touch one another. Table 5-1 shows what the data might look like. We now know that something distinctive happened on Thursday night and that on all the other nights there was roughly the same amount of touching. Would we know this if we had observed the group just on Wednesday and Thursday? No.

Suppose that the familiarity hypothesis were correct; the data might then look like those shown in Table 5-2. Since we have a whole series of observations, we are not likely to make the mistake of attributing the Wednesday-Thursday difference to the effects of alcohol.

Or suppose that the range of variability in touching is greater than we would have imagined and that the data looked like those shown in Table 5-3. In this case we would have to conclude that the amount of touching was controlled by extraneous variables. In all three sets of data, people touched one another 5 times on Wednesday and 15 times on Thursday; if we had measured only on Wednesday and Thursday, we would have no clue as to the role of alcohol in this change. By examining these data samples, however, it is evident that only

TABLE 5-2

Alcohol served	Day	Number of touches
No	Monday	0
No	Tuesday	1
No	Wednesday	5
Yes	Thursday	15
No	Friday	25
No	Saturday	40
No	Sunday	60

TABLE 5-3

Alcohol served	Day	Number of touches
No	Monday	8
No	Tuesday	21
No	Wednesday	5
Yes	Thursday	15
No	Friday	9
No	Saturday	2
No	Sunday	13

in the first case does our hypothesis about the effects of alcohol receive any support; thus, the advantages of the time-series design over the one-group pretest-posttest are apparent. Of course, we hasten to add that the time-series design does not automatically eliminate all the threats to internal validity inherent in the latter design. *History* is still a major threat to internal validity. That is, anything else that happened Thursday, before or during the group meeting, might have been responsible for the increase in touching. There might have been an article in the newspaper that day on the importance of expressing positive feelings; or Thursday might have been the only sunny day of the week. This problem, the investigator's inability to control adventitious events, is the primary disadvantage of this design. However, depending on patterns revealed in the data, multiple observations before and after the treatment enable researchers to use logic to rule out many of the other threats to internal validity.

Equivalent Time-Samples Design

Above we noted that history was the most serious threat to the internal validity of the time-series design when a single administration of a treatment is imbedded within a series of observations. We noted that it was possible that an event unrelated to the treatment could conceivably occur, resulting in an erroneous causal inference about the impact of the treatment. There is a partial solution to this problem. If the independent variable being studied is assumed not to have permanent effects on the behavior being observed, a slight modification of the previous design can add a large measure of control over undesirable influences from outside events. Instead of presenting the treatment once during the series of observations, we present it several times. Such a design—called an **equivalent time-samples design** (Campbell & Stanley, 1966, p. 43)—involves careful observations during a period in which the treatment is absent (a baseline or control period), followed by observations during a period in which the treatment is present (an experimental period), followed by another control period, another experimental period, and so on.

This modification is applicable to the alcohol study, since we expect the incidence of touching to return to a normal, baseline level after the influence of the wine has worn off. Applying this modification to the alcohol study, we would assign the treatment (alcohol) to days of the week, at random if possible. Thus, we might end up with the schedule shown in Table 5-4. If the hypothesis is confirmed, the data may look something like Table 5-5.

If the experiment is continued over a few weeks and treatments can be assigned to days at random, we can conclude that extraneous events are not causing the increase in touching. Even if we cannot assign treatments at random, we still have a great deal more confidence in the effect than we did when the treatment was administered only once. Any extraneous events that increased touching would have had to "just happen" to occur on the same days that we gave the group alcohol; therefore, the more days on which we conduct

TABLE 5-1

Alcohol served	Day	Number of touches
No	Monday	4
No	Tuesday	3
No	Wednesday	5
Yes	Thursday	15
No	Friday	7
No	Saturday	4
No	Sunday	5

posttest design would still remain. For example, the group members might be more relaxed the second evening, because the situation is more familiar to them. We are more confident about drawing conclusions from a time-series study, because this type of study allows us to see how distinctive the alcohol night is, compared to a number of similar occasions without the alcohol.

A specific example will allow us to explain this advantage more clearly. Suppose that one of the measures is the number of times people touch one another. Table 5-1 shows what the data might look like. We now know that something distinctive happened on Thursday night and that on all the other nights there was roughly the same amount of touching. Would we know this if we had observed the group just on Wednesday and Thursday? No.

Suppose that the familiarity hypothesis were correct; the data might then look like those shown in Table 5-2. Since we have a whole series of observations, we are not likely to make the mistake of attributing the Wednesday-Thursday difference to the effects of alcohol.

Or suppose that the range of variability in touching is greater than we would have imagined and that the data looked like those shown in Table 5-3. In this case we would have to conclude that the amount of touching was controlled by extraneous variables. In all three sets of data, people touched one another 5 times on Wednesday and 15 times on Thursday; if we had measured only on Wednesday and Thursday, we would have no clue as to the role of alcohol in this change. By examining these data samples, however, it is evident that only

TABLE 5-2

Alcohol served	Day	Number of touches
No	Monday	0
No	Tuesday	1
No	Wednesday	5
Yes	Thursday	15
No	Friday	25
No	Saturday	40
No	Sunday	60

TABLE 5-3

Alcohol served	Day	Number of touches
No	Monday	8
No	Tuesday	21
No	Wednesday	5
Yes	Thursday	15
No	Friday	9
No	Saturday	2
No	Sunday	13

in the first case does our hypothesis about the effects of alcohol receive any support; thus, the advantages of the time-series design over the one-group pretest-posttest are apparent. Of course, we hasten to add that the time-series design does not automatically eliminate all the threats to internal validity inherent in the latter design. *History* is still a major threat to internal validity. That is, anything else that happened Thursday, before or during the group meeting, might have been responsible for the increase in touching. There might have been an article in the newspaper that day on the importance of expressing positive feelings; or Thursday might have been the only sunny day of the week. This problem, the investigator's inability to control adventitious events, is the primary disadvantage of this design. However, depending on patterns revealed in the data, multiple observations before and after the treatment enable researchers to use logic to rule out many of the other threats to internal validity.

Equivalent Time-Samples Design

Above we noted that history was the most serious threat to the internal validity of the time-series design when a single administration of a treatment is imbedded within a series of observations. We noted that it was possible that an event unrelated to the treatment could conceivably occur, resulting in an erroneous causal inference about the impact of the treatment. There is a partial solution to this problem. If the independent variable being studied is assumed not to have permanent effects on the behavior being observed, a slight modification of the previous design can add a large measure of control over undesirable influences from outside events. Instead of presenting the treatment once during the series of observations, we present it several times. Such a design—called an **equivalent time-samples design** (Campbell & Stanley, 1966, p. 43)—involves careful observations during a period in which the treatment is absent (a baseline or control period), followed by observations during a period in which the treatment is present (an experimental period), followed by another control period, another experimental period, and so on.

This modification is applicable to the alcohol study, since we expect the incidence of touching to return to a normal, baseline level after the influence of the wine has worn off. Applying this modification to the alcohol study, we would assign the treatment (alcohol) to days of the week, at random if possible. Thus, we might end up with the schedule shown in Table 5-4. If the hypothesis is confirmed, the data may look something like Table 5-5.

If the experiment is continued over a few weeks and treatments can be assigned to days at random, we can conclude that extraneous events are not causing the increase in touching. Even if we cannot assign treatments at random, we still have a great deal more confidence in the effect than we did when the treatment was administered only once. Any extraneous events that increased touching would have had to "just happen" to occur on the same days that we gave the group alcohol; therefore, the more days on which we conduct

TABLE 5-4

Day	Alcohol served
Monday	No
Tuesday	No
Wednesday	Yes
Thursday	No
Friday	Yes
Saturday	No
Sunday	Yes

TABLE 5-5

Day	Alcohol served	Number of touches
Monday	No	3
Tuesday	No	4
Wednesday	Yes	16
Thursday	No	3
Friday	Yes	12
Saturday	No	6
Sunday	Yes	15

the study, the less likely this alternative hypothesis (history effects) becomes. It should be noted that this design is not feasible if the effects of the independent variable are not temporary. Moreover, even when treatment effects are not permanent, care should be exercised in choosing the interval between observations; enough time must elapse to enable subjects to return to their pretreatment states before control observations are made. On the whole, however—provided that proper precautions are taken to control potential sources of systematic error, such as the effects of testing, experimenter bias, and so on (see Chapters 8 and 9)—this is not a bad design, even though it lacks a separate control group.

Nonequivalent Control-Group Design

In previous sections we discussed two basic methodological attempts to gain a measure of control in a nonexperimental situation. The first method involves measuring the same group before and after the experimental treatment. This technique was the basic source of control in the one-group pretest-posttest design and in the two time-series designs. One of the major difficulties with this technique is that it fails to control for extraneous events; other incidents or changes between the pretest and the posttest may have affected the posttest observations. Of course, the equivalent time-samples design makes these other interpretations less plausible, but it can be used only with treatments that will "wear off" between sessions. The second type of control, used in the static-group comparison design, involves giving the treatment to one group and measuring two groups: the one that received the treatment and another group. The main problem with this design is that there is no way of telling whether the two groups were the same at the outset.

By combining these two types of design, we can create a study that substantially reduces both of these sources of error. Campbell and Stanley (1966, p. 47) call the combination design a **nonequivalent control-group design**. From the pretest-posttest design, we take the control provided by pretesting; from the static-group comparison design, we take the control provided by using two groups. Both groups are pretested, one group is given the treatment, and then

both groups are posttested. We choose groups that are as similar as possible and try to make our pretreatment measures simultaneous.

If you were to use this combination design in the alcohol study, you might arrange in advance for a friend to have a party on the same night as yours, to play the same sort of music, to invite the guests to come at the same time, to engage in approximately the same sort of activities, and so on. On the night of the party, you would conduct your observations (at both parties) as usual from 8 to 10 P.M.. Then, you would serve the wine to the guests at your own party only. From 11 P.M. to 1 A.M. you would continue your observations, again alternating between the two parties. Suppose the data on the number of touches at two different parties at two different times were as in Table 5-6.

This pattern of results tells us a good deal. First, since it is apparent that the groups were the same before you handed out the alcohol at your party, you have obtained the evidence of the baseline similarity that you wanted. Second, it tells us that the people at your party touched one another more after you handed out the alcohol: more than they had before, and more than the people at the other party did during the same period.

Many of the problems of the pseudoexperimental one-group pretest-posttest design (see pages 167–168) are now brought under control; if fatigue, familiarity with the group, or other temporal factors are causing changes in the guests' behavior, you will expect these effects to show up at both parties, and your data will be able to show this difference (see Table 5-7).

In the one-group pretest-posttest design you would have had only the data from the "your party" column and might have mistakenly attributed the effect to the alcohol. With the data on both parties before you, you would not be likely to make this error, since you can see that the same increase in touching occurred even without the alcohol. Likewise, though you cannot control the equivalence of the two groups before the introduction of the alcohol, you at least have the data from your pretest observations to help you decide whether it is reasonable to assume equivalence. If the groups are behaving differently between 8 P.M. and 10 P.M., your data will also show this difference (see Table 5-8).

In the static-group comparison design, your only data would be the post–11 P.M. observations, and you might decide that alcohol was making a difference. With the pretest and posttest data from both parties before you, you will not draw the false conclusion that the alcohol created the differences between the

TABLE 5-6

Observation times	Your party	Other party
8 P.M. to 10 P.M.	5	5
	(alcohol)	
11 P.M. to 1 A.M.	15	5

TABLE 5-7

Observation times	Your party	Other party
8 P.M. to 10 P.M.	5	5
	(alcohol)	
11 P.M. to 1 A.M.	15	15

TABLE 5-8

Observation times	Your party	Other party
8 P.M. to 10 P.M.	5 (alcohol)	0
11 P.M. to 1 A.M.	15	10

groups; you can see that the size of the differences (five touches per time period) was the same both before and after the alcohol.

The remaining serious source of error is the problem of other specific events that occurred at one party but not at the other. Any such event might be responsible for the posttest differences between the groups, and the investigator using this design must be continually alert to such alternative explanations.

Multiple-Group Time-Series Designs

By extension, it is readily apparent that increased control may be achieved by adding a second (or third) group to the time-series designs discussed above, especially to the time-series experiment (simple time-series design). Thus, if two groups met every day for a week, but only one group received alcohol on any given night, even the "special events" interpretation mentioned above would become less plausible, because it would be unlikely for some extraneous event to occur only on nights when one group was administered alcohol, and always to that particular group. You also gain a great deal of confidence in the results when you are able to give alcohol to one group on some nights and to the *other* group on other nights, because you can show that the results are not specific to your friends. If you can make the assignment of treatment random, you are within an ace of having an experimental study; the only thing that stands in your way is the random assignment of individuals to one or another group. Of course, in this example, the idea of getting the same group of people to come to seven parties at the same place in the same week may seem a little implausible; but there are many natural settings, such as classrooms and workplaces, in which regular group meetings are standard operating procedure. These kinds of situations are tailor-made for quasi-experiments such as **multiple-group time-series designs**. Not only does the researcher sacrifice relatively little of the control extant in a true experiment, but the question of generalizability of results to a real-world setting is moot.

GENERAL TECHNIQUES FOR IMPROVING NONEXPERIMENTAL DESIGNS

In following the alcohol study through its various developments, we have seen that the adequacy of the comparison group is one of the major design features

that determines our confidence in the results of a nonexperimental study. The other most important defense against error and uninterpretable results is the investigator's ability to introduce the treatment. For the sake of simplicity, we did not raise this as a central issue in the alcohol series; in all cases, we assumed that the experimenter was free to decide when and if to serve drinks to the group. The pitfall for an investigator who cannot control the occurrence of the treatment is readily apparent. The researcher does not know what causes a particular group to be exposed to the treatment (in this case, to drink alcohol); a group that is naturally exposed to a treatment may differ in a variety of treatment-related ways from one not exposed to the treatment. Any of the environmental or personality variables that result in a group's imbibing alcohol may also cause it to behave differently from a group that does not drink; any of these potential "third variables" can cause differences in our observations.

Each of the nonexperimental designs we presented can be conceptualized as having a relatively strong form and a relatively weak from, depending on whether the experimenter can *control the occurrence* of the variable under study. If this is not possible, the study is purely correlational. If it is possible, the researcher's study falls somewhere in between a correlational study and an experimental study. And, of course, if the investigator can assign the treatment to individual subjects at random, the study is an experiment.

To sum up, in designing a nonexperimental study, the investigator should concentrate first on trying to devise a situation in which the variables under study can be manipulated. In some cases this will be patently impossible; for example, social class, age, sex, IQ, and a large number of other variables cannot be manipulated. Other variables—such as teaching methods, working conditions, or advertising campaigns—can be manipulated, but cannot always be assigned to subjects at random. The second step is to determine, if possible, who will get the treatment, when it will be introduced, and what other group(s) will be observed. In all nonexperimental studies, even when variables cannot be manipulated, the investigator may improve the study by carefully considering whose behavior will be measured and when. Our confidence in the possibility of causation in the relationship grows in direct proportion to these improvements in the design.

The Importance of Reliable and Valid Measurement

Once the design has been chosen, a number of techniques can be employed during the actual running of the study to add to the experimenter's confidence that the relationship observed is a true one; most of these techniques involve procedures for making the observations themselves. Even a straight correlational study can be greatly improved—in terms of its ability to suggest specific hypotheses and variables worth following up—by careful attention to measurement.

Since correlational designs usually consist of nothing more than two sets of observations (measurements), the more specific and controlled the observations are, the clearer the results will be. In formulating a research question or

hypothesis, it is always a good idea for the researcher to think ahead to anticipate the possibilities for error. The investigator can control some of these by choosing one of the more "advanced" nonexperimental designs, by carefully selecting appropriate control groups, by finding a concrete, easily measurable behavior to represent a global conceptual variable, and by introducing other specific techniques to increase measurement standardization and decrease bias (see Chapters 8 and 9). Even if some types of error—such as possible third-variable correlations—cannot be controlled, the "informed" researcher is still in a much better position than the one who is ignorant of such potentially disruptive factors. Of course, it is best to have control of as many aspects of the situation as possible, but when this is impossible, awareness of exactly *which* aspects are uncontrolled is extremely important. The researcher can use this knowledge in a variety of valuable ways. First, awareness of other variables that may affect the behavior under observation may allow the researcher to measure these variables and analyze the data to determine whether any of them were confounded with the primary variable of interest.

For example, a researcher may hypothesize that people who talk a great deal are more likely to be perceived as leaders than are quiet people, but may also realize that many other variables—such as IQ, past leadership experience, and social class—can affect whether a person is perceived as a leader. Therefore, in addition to asking subjects to rate how good a leader a given person is and measuring the amount of time that person spends talking, the researcher may also give the "leader" an IQ test, ask the person (and others) about his or her leadership experience, and determine that person's social class. If people who are perceived as leaders talk more but are no different from nonleaders on the dimensions of IQ, experience, and class, the researcher has more confidence in the hypothesis; a certain amount of "control" has been achieved by *measurement* of possible third variables.

Anticipation of such third variables can also guide the investigator to the choice of a subject sample in which the effects of an extraneous variable are relatively improbable. For example, it has often been hypothesized that boys whose fathers were away from home during the child's first few years of life will turn out to be less "masculine" than boys whose fathers were present all along. However, since father absence is much more common in the lower classes than in the middle classes, the results of most studies are indissolubly confounded with the effects of social class. Aware of this problem, Kuckenberg (1963) tested the hypothesis in a sample of doctors' sons. Thus, the multitudinous extraneous variables associated with social class and occupation were ruled out as alternative interpretations of the results.

Many of the specific techniques discussed in the rest of this book as improvements for experimental studies can also be applied to nonexperimental studies. The fact that a situation does not allow an experimental study does not mean that rigorous methods should be abandoned. Techniques for eliminating bias, measuring the dependent variables, and avoiding problems common to social psychology experiments can all be exercised outside the laboratory. In many ways, a nonexperimental study demands greater creativity, since the obstacles to the achievement of meaningful results often increase as the researcher relinquishes control, but this should certainly not discourage the researcher from venturing outside the lab.

From the Field to the Laboratory to the Field (and Back)

In pointing out the need for creativity, we raise the possibility of going beyond existing techniques. The list of nonexperimental designs presented in this chapter is certainly not exhaustive; the designs are simply examples, and you may be able to think of better ones. Social psychology has been dominated by laboratory experiments for some 30 years, and there is a great need for innovative and informative nonexperimental research. Any given problem can be investigated in a variety of ways: as a laboratory experiment, as a field experiment, or as a nonexperimental study in either the lab or the field. By choosing any technique, the researcher invariably sacrifices the ability to collect clear data on some aspect of the problem, often because such a sacrifice is necessary to collect data on some *other* aspect. To choose a situation that best fits the problem, the investigator must decide what exactly is to be studied and what can be sacrificed for the sake of this information.

To cover a problem thoroughly, the investigator may decide on a series of different kinds of studies, so that the strengths of one sort of study compensate for the weaknesses of others, and vice versa. For example, early nonexperimental studies conducted in the field can provide ideas, hypotheses, and suggestions that might never occur to the experimenter sitting in an office and meditating over the question. These ideas and hypotheses can be refined, modified, and improved by careful follow-up studies (experimental and/or nonexperimental), and the effects of the basic variables suggested can be subjected to the rigorous test afforded by the experimental method.

On the basis of laboratory experimentation, cause-and-effect relationships can be sorted out. Some variables may be discarded as irrelevant; some may appear to be more powerful than they had previously. Various smaller aspects of the larger, real-world problems can be tested and confirmed or disconfirmed. New variables and hypotheses that seem more basic than the old ones may be discovered and tested. At this point, armed with these new refined hypotheses, the experimenter may do well to return to the field to see how the variables fare when they are thrown into the pot with other variables in a natural setting.

An excellent example of the value of moving from the laboratory to the field comes from the literature on the effects of mood on altruism. Techniques for inducing a positive or negative mood have been developed in laboratory settings. These mood inductions typically involve having subjects read affectively positive or affectively negative passages (e.g., Aderman, 1972) or having them reminisce about happy or sad experiences in their own past (e.g., Moore, Underwood, & Rosenhan, 1973). In the typical laboratory experiment, after

subjects have experienced the mood induction, they are given the opportunity to exercise their generosity by donating money for a good cause or by helping a confederate of the experimenter. Although results are mixed, studies employing this technique generally show that a positive mood increases helping. Despite many replications of this effect—across time, across laboratories, across investigators—the validity of these experiments has been questioned by some. Specifically, some critics challenge the artificiality of the setting in which helping behaviors are solicited and performed; others point to potential demand characteristics associated with the rather unusual mood-induction experience.

To counter these criticisms, researchers in the area have used field settings. Movie theaters are excellent settings for "natural" mood-induction treatments. That is, thousands of people attend movies: comedies that make them happy, or tragedies that make them sad. Benton Underwood and his colleagues (1977) took advantage of the emotional impact of motion pictures in a study of the effects of mood on helping. After careful pilot research, these investigators chose a double feature of *Lady Sings the Blues* and *The Sterile Cuckoo* as a (negative mood) treatment condition and chose two other double features to serve as neutral control conditions. The dependent variable (helping) was measured using a commonly occurring event: solicitation of donations to a nationally known charity, with collection boxes placed outside the movie theater lobby.

Of course, a major design problem encountered by Underwood and his colleagues (1977) was the fact that people do not assign themselves to movies at random. That is, *selection* was a potential threat to the internal validity of their study. If these researchers found that people who attend sad movies donated significantly less, they could not be sure whether it was something about *people who choose to attend sad movies*—and not the movies themselves—that accounted for the difference in helping. Although random assignment of movie goers to conditions was a logical possibility, there were drawbacks. For example, it would have been logistically difficult. And more important, problems of artificiality and reactivity—the very problems the use of field settings can prevent—would have been reintroduced. Therefore, the researchers decided to live with the selection threat and to alter the design to take it into account. To eliminate selection as a plausible rival hypothesis, they randomly alternated the timing of solicitation across different nights. That is, on some (randomly determined) nights, the opportunity to donate to the charity was available to people as they were entering the theater, *before* they had seen the double feature. On other (randomly determined) nights people had the opportunity to donate as they were leaving, *after* they had seen the double feature. Donations collected as people entered the theaters served as a check on the comparability of the groups before the treatment, that is, before they had seen a sad or neutral double feature. Fortunately for the investigators, the groups did not differ in their initial donation rates as a function of the movie they chose to attend. This pattern of results preserved the logic of random assignment—initial equivalence between experimental and control conditions—despite the procedural deviation from that ideal. Moreover, people who had viewed sad movies contributed significantly less on their way out of the theater than did people who had seen neutral movies. Thus, the results also supported conclusions based on "artificial" laboratory experiments.

We'd like to make two points related to this example. First, this study conducted in the field was not and could not be simply a "transplanted" replication of laboratory procedures. Major alterations were necessary to take advantage of the field setting. The researchers had considerably less control in the theater settings. They could not control administration of the treatment; neither could they control extraneous sources of variation. On any one night, a host of irrelevant events may have occurred during the course of the movies; projectors might have broken down or a disturbance might have erupted in the audience, and these extraneous events could have been confounded with the mood-induction treatment. The researchers not only were helpless to prevent such events but would not even have been aware of them had they occurred. In addition, as we've already mentioned, the experimenters were unable to assign subjects at random to conditions; in essence, they had to rely on "luck" to establish the initial equivalence among the groups.

Second, results of this quasi-experiment *as a single isolated study* would have been difficult to interpret without the context of conceptually similar laboratory experiments. This difficulty is partly due to the ambiguities introduced by altering the design of laboratory experiments and partly due to constraints inherent in a field setting where some degree of control is often sacrificed and where manipulation checks and random assignment, for example, are not always possible. The convergence of results across methods and across settings greatly enhances our confidence in both sets of findings. However, had the field study *failed* to replicate the laboratory results, numerous alternative explanations would have rendered interpretation very difficult.

Field studies—experimental and nonexperimental—are valuable in delimiting the parameters of the applicability of laboratory research. The researcher may find that the generality of some of the hypothesized causal laws are questionable, since in the real world other variables are always present to modify the effects isolated in the lab. Once these other variables are identified, the researcher may then return to experimental research, this time varying more variables at a time, testing new combinations and relationships among variables, refining hypotheses, building a theory. Natural settings can provide the evidence necessary to determine the generality of the results and to suggest new variables that must be brought under control if the research is to have widespread applicability, and experimental studies can be used to test causal relationships.

In summary, the right question is not: Are nonexperimental methods valuable? Every nonexperimental research method—correlational studies, pseudo-

xperiments, and quasi-experiments—has its place in the social psychologist's array of research tools. When conducted with ingenuity, careful thought and planning, and attention to measurement concerns, nonexperimental studies can serve as inspiration for theorizing *and* as means to evaluate and refine theory derived from experimental results. Thus, a far better question to ask is: Given my goals, given my research question, given methodological demands of the setting in which I choose to answer that question, which is the tool that best suits the requirements of the task?

CHAPTER 6

FINDING AND CREATING SETTINGS

In a typical sequence of events, the experimenter first has a general idea—a hypothesis, a notion of a relationship between two variables, or just curiosity—about the effects of one variable on another. Then, the experimenter must determine how to translate that idea into a social situation. The problem is not unique to social psychology, but many issues are more salient here than in other disciplines, because there are few standard procedures in experimental social psychology. Once experimenters have decided on the *form* the research question will take (i.e., the specific design to be employed), the next step is to decide on the *content* of the question (i.e., What setting will be used? What specific rationale will be provided to subjects? How will the independent variable be operationalized? What will constitute the dependent-variable measures?). There are no hard and fast rules about how this is done, and choosing and creating settings demands much in the way of imagination, ingenuity, and critical thought.

The experimenter starts with an idea or question about a causal relationship between two conceptual variables: Does lowered self-confidence make one more susceptible to the temptation to cheat? When someone asks a favor, is a person who is feeling guilty more likely to comply? Is aggression inhibited when the victim looks into the attacker's eyes? Whatever the question, the experimenter must create a workable set of procedures that address it fairly. This task demands not only the invention of a procedure for delivering the independent variable and measuring its effect on the subject's behavior but also the construction of a contextual framework, a *setting* in which the treatments make sense and have impact, the measurements are sensitive and accurate, and all elements of the experiment are plausible. In practice, these three problems of designing and conducting an experiment (treatment, measure, and set-