Experimental Designs

By now you should have some pretty clear ideas about the experimental approach—its key features, strengths and weaknesses, and the steps in carrying out an experiment. This chapter will provide additional knowledge necessary to become an informed consumer of research. Understanding the language and logic of experimental designs should be especially valuable to you as a citizen, as experimental designs increasingly are applied to the study of social policy issues. Reading this chapter is, in fact, a prerequisite to reading chapter 14, on the nature of policy assessment research.

Although the topic of experimental design may sound formidable and some experiments with their statistical baggage do appear rather complex, the basic principle of good design is simply the idea of "doing only one thing at a time." For the results of an experiment to be as unequivocal as possible, the only plausible explanation of changes in the dependent variable must be the manipulated independent variable. Therefore, a good design is one that rules out explanations of the results other than the effect of the independent variable. We learned in chapter 7 that this is best accomplished by randomly assigning subjects to experimental conditions (thereby controlling for pre-existing subject differences) and by making sure that the events occurring within each experimental condition are exactly the same except for the manipulated independent variable (thus controlling for extraneous factors and experiences during the experiment). The principle, then, is to allow only one factor, the independent variable, to vary while the rest are held constant.

In this chapter, the first three designs we consider are inadequate; they do not follow the principle of good design. Before we examine the flaws in these "pre-experimental" designs, we identify the kinds of uncontrolled variables, or threats to internal validity, that provide plausible rival explanations of study results. Next, we consider three basic, true experimental designs, which stand up better against threats to validity and thereby yield less ambiguous results. Then, we examine designs for testing the joint effects of two or more independent variables. The need for these more complex "factorial" designs exists because in the "real world," unlike in the typical laboratory experiment, many variables may be at work simultaneously. Finally, we examine some "quasi-experimental" designs, used when it is not practical or possible to meet all the conditions of true experiments.

Threats to Internal Validity

In chapter 7 we introduced the notion of internal validity. Recall that an experiment has internal validity when one can make strong inferences about cause and effect, inferring with confidence that the independent variable, rather than an extraneous variable, has produced the observed differences in the dependent variable. When we cannot separate the effects of the independent variable from possible effects of extraneous variables, we say that the effects are confounded. A subtle example of confounding is a taste-preference study sponsored by Pepsi-Cola (Huck and Sandler, 1979). In an early version of what became known as the "Pepsi Challenge," people were asked to taste two beverages, Pepsi and Coca-Cola, and indicate which they preferred. Ostensibly to prevent bias, the study sponsors concealed which beverage was Pepsi and which was Coke by marking the glasses with a letter, and perhaps to simplify the analysis, they used the same letters throughout the test: "M" for Pepsi and "Q" for Coke. Although Pepsi was reported to have won the Pepsi Challenge, the test lacked internal validity because the letter on the glass was confounded with the beverage in it. Indeed, when the study was replicated with Coke in both glasses, participants preferred the letter "M" over the letter "Q." Thus, letter preference rather than taste preference could have accounted for the original results.

The ideal research design effectively controls extraneous variables that threaten the internal validity of a study. We say "threaten" because uncontrolled extraneous variables pose explanations of study results that rival the hypothesized effects of the independent variable. To facilitate the evaluation of research designs, investigators have classified several common threats to internal validity, which we will now consider.

Each of these threats signifies a distinctive class of extraneous variables.

One threat to internal validity is history. This consists of events in the subjects' environment, other than the manipulated independent variable, that occur during the course of the experiment and that may affect the outcome. In this sense a "historical event" may be a major event of social or political importance, such as an assassination of a public figure or a prolonged strike, or it may be a minor event that occurs within the experimental setting and has no significance outside it, such as a hostile remark by a subject. For example, suppose that you are studying the impact of a series of written persuasive communications on attitudes toward U.S. immigration policies. During the course of your experiment, many of your subjects happen to view at home a feature television program on the plight of illegal immigrant workers. Should the results of your study show that subjects' attitudes toward U.S. immigration policies have changed, it would be impossible for you to tell whether this change was caused by your independent variable (written communications) or by the television program. The confounded effects of the experimental manipulation and history preclude a clear causal interpretation of the results.

Another frequent rival explanation for research findings is maturation. By this we mean any psychological or physical changes taking place within subjects that occur with the passing of time regardless of the experimental manipulation. Even during a 1- or 2-hour experiment, subjects may become hungry or tired. Over a long-term experiment, subjects may grow physically or intellectually, become more rigid or more tolerant, or develop health problems or improved health. The effects of such maturational factors may be confounded with treatment effects. For example,
in a study of the effectiveness of a new physical therapy program for stroke victims, any progress that the stroke victims would make naturally over time without therapeutic intervention (a phenomenon known as "spontaneous remission") might incorrectly be attributed to the therapy.

**Testing** represents a third possible source of internal invalidity. Similar to reactive measurement effects discussed in chapters 5 and 7, testing refers to changes in what is being measured that are brought about by reactions to the process of measurement. Typically, people will score better or give more socially desirable or psychologically healthier responses the second time a test or scale is administered to them. This is true even when a different but similar measure is used. There are a number of reasons this occurs. On some measures, such as intelligence tests, the tasks simply become easier after practice. Attitude scales, on the other hand, may alert subjects to the purpose of the scale, causing them to give socially desirable responses or perhaps to re-examine their own attitudes. Such effects are potentially confounded with the effects of the independent variable whenever subjects are measured twice in the same study and the initial measurement arouses their awareness of being studied.

A fourth threat to internal validity, **instrumentation**, refers to unwanted changes in characteristics of the measuring instrument or in the measurement procedure. This threat is most likely to occur in experiments when the "instrument" is a human observer, who may become more skilled, more bored, or more or less less observant during the course of the study. Instrumentation effects also may occur when different observers are used to obtain measurements in different conditions or parts of an experiment. Such an effect would be analogous to a shift in an instructor's grading standards while grading a set of essays or to inconsistent standards on the part of two instructors grading subsets of the same essays.

The fifth threat to internal validity is **statistical regression**, the tendency for extreme scorers on a test to move (regress) closer to the mean or average score on a second administration of the test. Also known as **regression toward the mean**, this phenomenon is likely to affect experimental results when subjects are selected for an experimental condition because of their extreme scores. Consider, for example, an experiment to assess the effectiveness of an assertiveness-training program in helping shy, introverted persons become more socially outgoing. Subjects are given a previously validated scale of extraversion, and the most introverted subjects—those who score in the bottom quartile—are assigned to the experimental condition, where they undergo an 8-week assertiveness-training program. The remaining subjects serve as a control group, receiving no such training. At the end of 8 weeks, both groups are retested. Assuming that subjects in the experimental group showed more significant gains in extraversion, on average, than the control group, we cannot confidently attribute this difference to the assertiveness-training program. The students chosen for assertiveness training were those scoring in the bottom 25 percent on a measure of extraversion. Because of this, the observed increase in extraversion scores may have been due to statistical regression rather than the training program.

Statistical regression can explain many events in everyday life. An example should clarify how its effects occur. Suppose that Instructor Young gives his class two exams. In each case the average grade is B. Table 8.1 compares the performance of individuals on the two tests. Young is disappointed to see, from the first row of Table 8.1, that of the fifteen persons who received As on Exam I, only ten received As on Exam II, with four receiving Bs and one receiving a C. He concludes that five of the students became overconfident after the first exam and slacked off. Then he looks at the C scorers and is delighted to see that of the fifteen who scored C initially, two raised their scores to A and four to B. He feels some satisfaction that at least these students hit the books to bring their grades up. Such conclusions could be correct, but it is more likely that the results are due to regression toward the mean.

Because measurement error is always present, there is never a perfect correlation between scores from separate administrations of the same test or measure. A relatively large amount of measurement error will be reflected in a low correlation of scores and will result in more regression toward the mean. Another way to think of it is that the extreme scorers, as a subgroup, are affected more by chance factors we might call luck. On the first exam, some of the high scorers are particularly lucky, whereas some of the low scorers are particularly unlucky. But it would be rare for extremely lucky students on the first exam to repeat their good fortune on the second exam (or for unlucky students to repeat their misfortune), which results in subgroup scores closer to the mean. Initially average scorers, on the other hand, are likely to include about as many lucky as unlucky individuals so that changes in luck will tend to cancel out and not influence results on the second test.

A sixth threat to internal validity, **selection**, is present whenever there are systematic differences in the composition of the control and experimental groups. Such selection bias is especially likely when naturally existing groups are studied. Suppose, for example, a study compared 1-year recidivism rates of apparently recovered alcoholics who received help through Alcoholics Anonymous with the recidivism rates of those who received in-hospital treatment. One possible systematic difference in the two groups that might affect abstinence is economic standing. If economic well-being aids recovery and if those choosing in-hospital treatment were in fact more affluent, then the hospital treatment group might have a better record of recovery because of their economic status, apart from any benefits of the treatment. Or, another possibility, also a matter of selection, is that the people entering the two treatment programs differed in the severity of their alcohol abuse, with more severe cases, who were also less likely to be "cured," receiving in-hospital treatment.
An additional example where selection would be a confounding factor is a study comparing the academic progress of pupils in an alternative school with that of pupils in a conventional school. Even if attempts were made to match the alternative school with a conventional school on such relevant characteristics as class size and socioeconomic status of pupils, there still might be confounding differences between the two groups. For example, the parents of the alternative-school children might be more permissive or differ in other unknown ways from the parents of the conventional-school pupils. Whenever the groups are not equivalent at the beginning of the experiment, it is difficult to interpret differences on the dependent variable.

Another rival explanation for experimental results involves the loss of subjects from the experimental groups. The reasons for subjects dropping out range from illness or moving out of the area to disenchantment with the experience of being a subject. The loss of subjects in an experiment is called attrition. Attrition poses the greatest threat to internal validity when there is differential attrition—that is, when the conditions of an experiment have different dropout rates. Invariably, those subjects who drop out differ in important ways from the ones who remain so that the experimental conditions are no longer equivalent in composition; thus, differential attrition undermines the effects of random assignment. In an experiment to test the effectiveness of certain behavior-modification techniques on fingernail biting, for example, nail biters might be randomly assigned either to the experimental treatment or to the control group. But what if nail biters who are making the least progress drop out of the experimental group? The measure of nail biting at the end of the treatment will tend to reflect greater success than if all subjects had remained in the experimental group. Thus, the effect of differential attrition will be confounded with the treatment.

Finally, experimental findings may be confounded by an interaction between two or more of the threats to internal validity we have discussed. This simply means that two of these threats, say selection and maturation, act together to affect the outcome on the dependent measure. You will recall the example we used to illustrate the possible confounding effects of maturational processes, a hypothetical study of the effectiveness of a new physical therapy program for stroke victims. It was pointed out that some degree of spontaneous remission might be confounded with the therapy effects. Now suppose that patients were allowed to select either the new therapy program, where treatment took place at a hospital, or their usual program, which involved visits by a physical therapist to the patient's home. We can conjecture how this process of selection might interact with effects of the normal recovery process (maturation). Those patients who volunteered for the new treatment program would differ from those who did not in one obvious way: They would have to have transportation to the hospital. For most patients, this would mean a family member or some other person available and willing to take them and wait for them while they had the therapy. Perhaps stroke victims who have someone around who is willing to make such an effort are also receiving more social and emotional support (and perhaps other benefits) than are those patients who do not have such a person available. Such benefits could possibly facilitate spontaneous remission for patients in this therapy group. Thus, the interaction effects of selection (differences between patients choosing and not choosing the new program) and maturation (spontaneous remission) would be confounded in this study with the effects of the therapy program.

We now examine three research designs in which threats to internal validity pose rival explanations for the results, thus preventing meaningful interpretation. Because these designs lack one or more features of true experimental designs, they are termed "pre-experimental."

**Pre-experimental Designs**

**Design 1: The One-Shot Case Study**

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>O</th>
</tr>
</thead>
</table>

In the simplest possible design, dubbed the one-shot case study, some treatment is administered to a group, after which the group is observed or tested to determine the treatment effects. The above diagram illustrates this design: X stands for the treatment condition of the independent variable and O stands for the observation or measurement of the dependent variable. Time moves from left to right.

To illustrate this design, imagine a teacher who is having a problem with her 4th-grade pupils frequently talking out of turn. She decides to conduct the following "experiment" one day. Every time a child speaks out of turn, the teacher says, "You are talking out of turn," and then immediately turns her attention to the pupil who was interrupted or to another who is modeling desirable behavior. (This is the treatment, symbolized by the letter X in the diagram.) Toward the end of the day a teacher's aide carefully records the number of incidents for each child and finds that few children are talking out of turn and that the number of incidents is low. (This is the measurement of the dependent variable, symbolized by the letter O.) The teacher concludes that her treatment is effective and recommends it to other teachers.

Unfortunately for the teacher's efforts, such a conclusion is clearly unwarranted; several other explanations may account for the apparent change in behavior. Maybe one of the more recalcitrant talkers left after lunch for a date with the dentist; this would result in an attrition effect. Maturational variables also might threaten the study's internal validity if, for example, some of the children were particularly tired that day. Finally, history could offer an alternative explanation for the results; perhaps early in the day the class had a music lesson in which they let off steam by doing a lot of exertive singing and so felt less inclined to interrupt during class discussions. Attrition, maturation, and history represent threats to the internal validity of any study based on the one-shot case study design.

The critical flaw in this design is that it provides no adequate basis for comparing the findings with other observations; and some process of comparison is essential to scientific inference. We cannot tell in our example what the incidence of talking out of turn would have been with a different intervention, with no intervention, with a different group of students, or on a different day. All we know is that the teacher intuitively sensed that talking out of turn decreased.
Many of our day-to-day assumptions about causality are based on "experiments" similar to one-shot case studies. Suppose a jogger buys a new brand of running shoes and afterward finds that she is running faster than before. Concluding that the shoes have "helped" her, she recommends them to you. Can you think of other explanations for her increase in speed?

**Design 2: The One-Group Pretest–Posttest Design**

\[ O_1 \quad X \quad O_2 \]

A second pre-experimental design, the one-group pretest–posttest design, involves observing or measuring a group of subjects (the pretest), introducing a treatment (the independent variable), and observing the subjects again (the posttest). The pretreatment observations are represented by \( O_1 \), the independent variable by \( X \), and the posttreatment observations by \( O_2 \). For example, the performance of a group of joggers might be timed before \( (O_1) \) and after \( (O_2) \) they received a new brand of running shoes \( (X) \). Design 2 is commonly found in educational, organizational, and clinical research. It is an improvement over design 1 because it provides a basis of comparison, but it is still subject to major sources of invalidity.

We may illustrate this design by changing our design 1 example slightly. Suppose the teacher's aide counted talking-out-of-turn incidents on the day before as well as the day of the experiment. How would this change in design affect the study's internal validity? First, the **threat of history** confounding the findings is controlled, for the most part, as the two groups should experience the same major environmental events. Since there is no pretest, the threats of testing and statistical regression are absent. And as long as measurements are equally reliable and valid for the two groups, instrumentation is not a problem. On the other hand, **selection** is a serious threat to internal validity, for without the random assignment of subjects to the experimental and control groups, there is no control of possible pretreatment differences.

**Experimental Designs**

A third pre-experimental design, the static group comparison, like design 2, is an improvement over the one-shot case study in that it provides a set of data with which to compare the posttreatment scores. Whereas design 2 provided pretreatment scores on the same group, design 3 provides the scores of a control group. As symbolized above, the rows represent separate groups: \( X \) stands for the experimental treatment, the blank space under \( X \) stands for the no-treatment control, and \( O \) represents the dependent-variable measure. Notice that each group is measured just once.

Our classroom example again may be altered slightly to fit this design. Let us suppose that, because of rapid growth in the area, the school is on double sessions so that the teacher has one class in the morning and a different one in the afternoon. On the treatment day she tries out her new approach on the morning class only but records talking-out-of-turn incidents for the pupils in both groups.

Although the static group comparison does a better job of controlling threats to internal validity than do the other two pre-experimental designs, some threats remain. The **threat of history**, confounding the findings is controlled, for the most part, as the two groups should experience the same major environmental events. Since there is no pretest, the threats of testing and statistical regression are absent. And as long as measurements are equally reliable and valid for the two groups, instrumentation is not a problem. On the other hand, **selection** is a serious threat to internal validity, for without the random assignment of subjects to the experimental and control groups, there is no control of possible pretreatment differences. 

**KEY POINT**

Specific threats to valid causal inference depend on the research design: Testing, history, and maturation are the main threats in a one-group pretest–posttest design; selection is the main threat with the static group comparison.
True Experimental Designs

The designs described in this section differ from the pre-experimental designs in that there are always two or more groups and subjects are assigned to them randomly to ensure approximate equivalence of the groups.

**Design 4: The Pretest–Posttest Control Group Design**

$$R \quad O_1 \quad X \quad O_2$$

$$O_1 \quad O_2$$

The pretest–posttest control group design involves measuring the experimental group before and after the experimental treatment. A control group is measured at the same time but does not receive the experimental treatment. As symbolized before, the rows represent separate groups, and time moves from left to right: the R to the left indicates that subjects are randomly assigned to the groups, each O stands for an observation, the X symbolizes the treatment condition of the independent variable, and the blank space under the X indicates the no-treatment control condition.

Our earlier example of the study of assertiveness-training effectiveness may be altered to illustrate this design. In the original example, which was not a true experiment, subjects who scored in the bottom 25 percent on an extraversion scale were assigned to the experimental treatment, with the remaining 75 percent serving as controls. To fit design 4, the entire pool of available subjects would be randomly assigned to the treatment and control groups. Both groups then would be given the extraversion scale (O1 and O2 pretest measurements). Only the experimental treatment group would receive the assertiveness training; at the conclusion of the training, both groups would again be given the extraversion scale (O2 and O4 posttests).

How does this design deal effectively with the common threats to internal validity? To consider history, context, any event in the general environment that would produce a difference between the pretest and posttest in the experimental group (O1–O2) would produce about the same difference in the control group (O1–O4). Similarly, changes due to maturation, testing, or instrumentation would be felt equally in both groups. Therefore, these factors cannot account for differences between the posttests, O4 and O1. Random assignment also eliminates the factors of selection and regression, within the limits of chance error. Comparison of O1 and O2 provides a check on the randomization procedure with regard to initial differences on the dependent variable. And even if the subject pool consisted only of extreme scorers—for example, all introverts—random assignment of these subjects to experimental and control groups should ensure initially equivalent groups that regress about the same amount on the posttest. Finally, this design permits the assessment of possible attrition effects; one can compare both the number of subjects and the pretest scores of those who drop out of each group.

Since true experimental designs adequately control threats to internal validity (without which we cannot tell whether the independent variable was responsible for the results), it is appropriate to examine these designs for possible threats to external validity. You will recall that external validity refers to the extent to which a study’s findings have meaning outside the particular circumstances of the experiment—that is, the extent to which the results may be generalized.

The pretest–posttest control group design suffers from the external-validity threat of testing interacting with the independent variable, called testing–X interaction or testing–treatment interaction. This simply means that the effect of the independent variable may be different when a pretest is present from when it is not.

**KEY POINT**

In a pretest–posttest control group design, the effect of the treatment may depend on the presence of a pretest. Known as testing–treatment interaction, this limits external validity.

Sometimes an independent-variable effect can be produced only with subjects who have been sensitized to the experimental treatment by pretesting. To continue with our example, it may be that the assertiveness training is effective in helping people become more socially outgoing only when they have been made particularly conscious of their introversion (or extraversion) by responding to the pretest extroversion scale. If that were true, results of the study could be generalized only to other similarly pretested groups.

The extent to which one need be concerned about an interaction between testing and treatment depends on the experimental situation. In educational settings, where test taking is the norm, the effects of a testing–treatment interaction would probably be negligible; and the learning situations to which one would be generalizing are likely to involve testing. The classroom experiment described in chapter 7 on the effects of teacher comments on student achievement (Page, 1958) is an example of the usefulness of the pretest–posttest control group design. Since regular classroom tests and procedures were used in this study, there is no reason to believe the pretests interacted with the treatment to any significant extent. On the other hand, in studies of attitude change or persuasion, a pretest may very well alert subjects to the treatment to follow in such a way as to make them more receptive (or resistant) to it. In such cases, the findings would have little external validity, and it would be better to use the following design.

**Design 5: The Posttest-Only Control Group Design**

$$X \quad O_1$$

$$O_2$$

The simplest of the true experimental designs, the posttest-only control group design incorporates just the basic elements of experimental design: random assignment of subjects to treatment and control groups, introduction of the independent variable to the treatment group, and a posttreatment measure of the dependent variable for both groups. Notice that, except for one crucial difference—subject randomization—design 5 resembles the pre-experimental static group comparison design. Unlike this design, however, design 5 controls for the common threats to internal validity adequately.
Some researchers seem to feel more confident that groups are equivalent before the experimental manipulation when they can check pretest scores. Yet, in reality, the random assignment of subjects is sufficient to ensure approximate equivalence. Therefore, under most circumstances, design 5 is preferable to design 4, the pretest-posttest control group design. By eliminating the pretesting step, design 5 has two major advantages. First, it is more economical. Second, and more important, it eliminates the possibility of an interaction between the pretest and the experimental manipulation. Still, there are special situations requiring a pretest and other situations in which a pretest would be useful. Imagine, for example, a long-term experiment in which you expect a higher than usual number of subjects to drop out. In such a situation, pretest scores on each group would help in determining if there was an interaction between attrition and the experimental manipulation.

**Design 6: The Solomon Four-Group Design**

\[
\begin{array}{cccc}
O_1 & X & O_2 & \\
R & O_3 & O_4 & \\
X & O_5 & O_6 & \\
\end{array}
\]

A third true experimental design, the Solomon four-group design is really a combination of designs 4 and 5, as may be seen in the above symbolic representation. Here we have an experimental group and a control group that are pretested, along with an experimental and a control group that are not pretested.

The Solomon four-group design has the advantages of both of the two previously discussed experimental designs; information is available regarding the effect of the independent variable (O₂ and O₅ compared with O₃ and O₆), the effect of pretesting alone (O₃ versus O₄), the possible interaction of pretesting and treatment (O₂ versus O₅), and the effectiveness of the randomization procedure (O₁ versus O₂). While this design provides more information than either of the other two experimental designs, the requirement of two extra groups makes it much more expensive to use.

**Within-Subjects Designs**

Each of the three true experimental designs requires the random assignment of subjects to conditions. Randomization is, of course, intended to create groups that are equivalent on uncontrolled extraneous variables. Another solution to the problem of equivalence is to have the same subjects participate in both the treatment and control conditions of the experiment. In other words, an experimenter might expose subjects to a treatment (X₁), apply the measure of the dependent variable (O₁), expose the same subjects to a control (comparison treatment) condition (X₂), and then measure the effect a second time (O₂), as diagrammed below:

\[
\begin{align*}
X_1 & \rightarrow O_1 \\
R & \rightarrow O_3 \\
X_2 & \rightarrow O_5 \\
\end{align*}
\]

This type of study is called a within-subjects design because each subject acts as his or her own control—that is, O₁ is compared with O₂ for each subject. By contrast, designs 4–6 are called between-subjects designs because different groups of subjects are compared with one another.

Within-subjects designs have a long history in the natural sciences and are used in some areas of psychological research such as sensation and perception. They have two principal advantages. First, they require fewer subjects; a within-subjects design with a treatment and a control condition, for example, would require half as many subjects as design 5. Second, by having each individual experience and react to every condition of the experiment, they reduce the error associated with how different people react to these conditions. But, alas, within-subjects designs also create threats to internal validity that often make them inappropriate for social research. A study by R. Brent Galuppo and associates (1992) illustrates these threats and the general strategy for estimating their effects.

Galuppo and colleagues sought to compare the effectiveness of traditional brainstorming with electronic brainstorming. From the time it was introduced in the 1950s, brainstorming has been a popular method for enhancing group creativity. To stimulate ideas through group discussion, members of brainstorming groups are asked to express as many ideas as they possibly can while withholding comments and criticisms. Despite its widespread use, however, research has shown that brainstorming groups do not outperform noninteracting individuals. Group researchers attribute this to various inhibitors, such as evaluation apprehension, present in interacting groups. To reduce these inhibitors, they recently have introduced a new technique called “electronic brainstorming” in which group members simultaneously type ideas into a computer, which then distributes the ideas to all members of the group.

To compare the productivity of traditional verbal brainstorming with electronic brainstorming in a within-subjects design, one simply could have the same groups generate ideas first verbally and then electronically. This is, in part, what Galuppo and colleagues did, and they used the number of nonredundant ideas each group produced as their primary measure of the dependent variable. Suppose more ideas were produced following electronic brainstorming than verbal brainstorming. Besides the different form of communication, what might account for this effect? One possibility is a testing effect; with practice, groups became better at generating ideas. (Alternatively, if they produced fewer ideas over time, this might be due to fatigue.) Another possibility is that subjects may become aware of the experimenter’s hypothesis; if subjects realized after the first, traditional brainstorming session that their idea generation was being compared under two conditions, they might behave differently in the second session, perhaps viewing electronic communication as a novel and more interesting challenge.

More generally, in any within-subjects design, there is always the possibility that observed changes are due to the sequencing or order of the treatment and control conditions rather than to the treatment. The principal method of controlling for order effects is counterbalancing. This consists of reversing the sequence of the
treatment and control conditions so that different groups of subjects experience either sequence. Gallup and colleagues' experiment was counterbalanced: half the subjects brainstormed verbally first and the other half brainstormed electronically first, with the order determined randomly. With random assignment, the within-subjects design becomes a true experiment, effectively controlling for threats to internal validity. Testing effects should be manifested equally in both sequences, and the effects of the order of the treatment and control condition can be estimated by comparing differences in the measures of the dependent variable in one sequence with differences in the other. But even though testing is manifested equally and order effects can be tested, the randomized within-subjects design does not eliminate these effects. Consequently, if it is likely that participating in one condition of an experiment will influence how subjects respond to another, as often occurs in social psychological experiments, this design should not be used.

Gallup and colleagues actually carried out two experiments comparing verbal with electronic brainstorming. Besides counterbalancing the different forms of communication, both experiments manipulated group size or the number of group members and both used two different idea-generation problems to reduce the testing effect. The researchers found that neither the type of problem nor the order of the brainstorming techniques had an effect, which simplified the interpretation of the findings. Except in two-member groups, electronic brainstorming produced more ideas than verbal brainstorming.

Overview of True Experimental Designs

Many simple variations of experimental designs are possible. Although designs 4 and 5 were presented as having just two groups (an experimental group and a control group), the logic of either design may be extended easily to three or more groups: One might, for example, want to compare several clinical approaches for treating depression or a number of methods of teaching reading. Similarly, groups may be added to vary the intensity of the independent variable; for example, we might induce a high level of frustration in one group, a moderate level in another, and a low level in a third group.

Sometimes ethical considerations preclude withholding treatment from a control group. This is frequently the case in the fields of correction, clinical psychology, medicine, and education. Also, a true no-treatment control group, one that is exactly identical to the treatment group except for the treatment manipulation, is impossible to implement in many situations. The hypothetical study discussed earlier, in which a new approach to physical therapy for stroke victims was compared with a standard approach, illustrates the common variation of experimental design in which there is no true control group; rather, two or more treatments are compared for their relative effectiveness.

We have pointed out the importance of random assignment of subjects as a means of controlling for pre-existing differences (known and unknown) in subjects. However, randomization introduces one other threat to internal validity: The observed results might have occurred by chance rather than being caused by the experimental variable. Recall that tests of statistical significance are used to determine the likelihood of this occurring, thereby screening out trivial results that could have occurred easily by chance. Having noted this, we stress again that the random assignment of subjects is an integral part of any true experimental design.

The matter of external validity warrants further comment. As noted earlier, external invalidity results from an interaction of the treatment, or independent-variable manipulation, with some other variable. The presence of such interaction means that treatment effects apply only under certain conditions inherent in the experiment. Earlier, for example, we discussed the threat of a testing–treatment interaction as a particular concern in the pretest–posttest control group design. This interaction limits the generalizability of results to situations in which subjects have been pretrained. External validity similarly may be threatened by interactions of the treatment with characteristics of the subject population, time, or some other feature of the experimental setting. We now elaborate on a few of these sources of external invalidity.

Sample selection often restricts external validity. Because the sample of subjects participating in an experiment typically consists of homogeneous groups such as college students, the possibility of sample selection interacting with the independent variable may be present in any of the experimental designs. Thus, extreme caution must be used in generalizing any effect of the independent variable to dissimilar groups. A cigarette smoking cessation treatment, for example, might succeed with volunteer subjects (who are highly motivated) but be ineffective with the general population of smokers who wish to quit.

Maturation also may interact with the independent variable; that is, the effect of the treatment may occur only with subjects in a certain physical or mental state. For example, the findings of an experiment conducted at four o’clock on a hot summer day may be generalizable only to hot, tired subjects. Solutions to this problem include (1) deliberately varying the conditions that would seem to affect maturation states as part of the experimental design and (2) replicating the basic experiment under varying conditions.

Finally, the effect of the treatment might be peculiar to the historical circumstances surrounding the experiment. For example, a study of the effects of certain experiences on Christians’ attitudes toward Jews might produce significant results only because the experiment took place at the time of a highly rated television series on the Holocaust. In viewing it, subjects became more receptive to the treatment. Replicating experiments under different historical circumstances is an effective way of ruling out the threat of such treatment-history interactions.

Factorial Experimental Designs

Social events often are caused or influenced by a number of variables. Therefore, it frequently makes sense to study several possible causes, or independent variables, at
METHODS OF DATA COLLECTION

the same time. When two or more independent variables are studied in a single experiment, they are referred to as factors, and the designs that enable us to explore their effects jointly are called factorial designs. Although more than one variable is manipulated, it is possible to assess the effect of any manipulated variable while controlling for the impact of other variables. Hence, the basic principle of good design, "doing only one thing at a time," still applies.

An interesting example of a factorial design is Jody Gottlieb and Charles Carver’s (1980) experiment on "the anticipation of future interaction and the bystander effect." As you may recall, Darley and Latané (1968) demonstrated the bystander effect: the larger the number of bystanders who witness an emergency, the less likely that any one person will help—in an experiment described in Chapter 7. Gottlieb and Carver’s study was quite similar to Darley and Latané’s. Subjects (college students) were asked to engage in a discussion of the problems of college living with other subjects over an intercom system; the other discussants were taped recorded; the number of discussants varied, either one or five others; and the dependent measure consisted of the length of time that subjects took to respond to a staged emergency (one of the supposed fellow discussants, who apparently had been eating, began choking and struggling for breath and cried out for help). So, one independent variable or factor was the number of discussants or "bystanders." Gottlieb and Carver reasoned, however, that the bystander effect may occur because of the anonymity of the subjects; that is, they never saw nor did they expect ever to meet their co-subjects face-to-face. If subjects expected to interact with other bystanders face-to-face in the future, they might be more willing to help to avoid being blamed by others for their inaction. Thus, Gottlieb and Carver manipulated a second factor, anticipated interaction, by telling some subjects that they would be meeting afterward face-to-face and telling the others that they would not meet the co-subjects at all.

The Gottlieb and Carver experiment represents the simplest case of a factorial design. Two variables (number of bystanders and anticipated interaction) are manipulated, with each variable having two levels or categories. This design requires four experimental conditions, one for each combination of variable categories: one discussant (the emergency victim)/anticipated interaction; one discussant/no anticipated interaction; five discussants/anticipated interaction; and five discussants/no anticipated interaction. When a design has two independent variables, each having two levels, we call it a $2 \times 2$ (two by two) factorial design. A design that had three levels of one factor and four levels of another would be a $3 \times 4$ factorial design having twelve conditions. In every case the number of conditions of a factorial design may be determined by multiplying the number of levels of the first factor by the number of levels of the second factor and, if there are additional factors, by the number of levels of each in turn. Theoretically, a factorial design may utilize any number of factors, although there are practical limits.

As you can see, factorial designs are simple extensions of the basic experimental designs. In fact, we already have introduced one design, the Solomon four-group design, which may be viewed as a factorial design. We depicted this design as follows, with each row representing a separate group.

<table>
<thead>
<tr>
<th>Experimental Designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1$ $X$ $O_2$ first group</td>
</tr>
<tr>
<td>$O_3$ $O_4$ second group</td>
</tr>
<tr>
<td>$X$ $O_3$ third group</td>
</tr>
<tr>
<td>$O_4$ fourth group</td>
</tr>
</tbody>
</table>

Table 8.2 presents the Solomon four-group design in factorial form. Note that this design has two factors: the treatment and the pretest. Also, like the Gottlieb and Carver experiment, both levels (categories) of one factor are combined with both levels of the second, forming the four experimental groups in the cells of the table. So, this is another example of a $2 \times 2$ factorial design.

Ideally, each cell will have the same number of subjects. When this is not possible (because of subject attrition or other reasons), the dependent-variable data in the various cells can be analyzed after statistical adjustment has been made to compensate for the uneven number of subjects. As always with true experimental designs, subjects are assigned to the various conditions (cells) by a random device to control for pre-experimental differences.

You may be wondering how the effects of the various factors are determined in a factorial design. To answer this question, let us consider a hypothetical study of the effect of a sympathetic movie portrayal of the gay community (the treatment) on attitudes toward gay rights (the dependent variable). The results from the Solomon four-group experiment are shown in Table 8.3. Each cell contains the average (mean) posttest attitude score for that experimental group. A higher score indicates a more positive attitude toward gay rights.

First, factorial designs provide information about the main effect of each factor—that is, the overall effect of the factor by itself. We have labeled the treatment variable "factor A" and the pretest variable "factor B." The main effect of factor A is determined by comparing the overall mean score of subjects who received the experimental treatment (which in this case is 25) with the overall mean score of subjects who did not receive the experimental treatment (in this case 15). Clearly, subjects who saw the movie were more supportive of gay rights than those not exposed to the movie. Whether this treatment effect is "significant" would have to be determined by an appropriate statistical test. Let us assume that all effects are statistically significant.

<table>
<thead>
<tr>
<th>Table 8.2. Solomon Four Group Design Represented as a $2 \times 2$ Factorial Design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment condition</strong></td>
</tr>
<tr>
<td>(factor A)</td>
</tr>
<tr>
<td>Treat</td>
</tr>
<tr>
<td>First group, $O_2$</td>
</tr>
<tr>
<td>Second group, $O_4$</td>
</tr>
<tr>
<td>Third group, $O_3$</td>
</tr>
<tr>
<td>Fourth group, $O_4$</td>
</tr>
</tbody>
</table>
Likewise, the main effect of factor B is determined by comparing the overall mean score of pretested subjects (30) with the overall mean of subjects not pretested (10). Thus, exposure to the pretest also enhanced posttest attitudes toward gay rights. While this result is not unexpected, it is nonetheless discouraging since ideally treatment effects (the movie) should be much more powerful than measurement artifacts (pretesting effects).

**Interaction Effects**

Notice that we determined the main effects of each factor by averaging over all levels of the other. Assessing main effects in this way actually can be very misleading as it may conceal how the two factors, acting together, produce the outcome. A major advantage of factorial designs is that they also provide information about the joint effects of the factors. If there is an interaction between two factors, the effect of one factor on the dependent variable varies according to the value or level of the other. In other words, the effects of the factors together differ from the effects of either alone. Because of this possibility, it is always best to check first for an interaction; otherwise, conclusions based on main effects may be incorrect.

The presence of the pretesting effect in our example should lead us to question the generalizability of the movie effect. Specifically, there may be an interaction between the pretest and the movie, with the pretest sensitizing subjects to the topic and thereby enhancing the movie’s impact. Inspection of the cell means in Table 8.3, graphed in Figure 8.1, does indeed reveal a testing-treatment interaction: The movie has an impact only on pretested subjects. Consequently, its value outside the laboratory as an attitudinal change agent appears to be very limited.

**Interaction effects** do not always look like the idealized results in Figure 8.1. Some other possible outcomes are graphed in Figure 8.2. The movie treatment may affect all subjects but more so for those pretested (Figure 8.2A), may have no effect on pretested subjects (Figure 8.2B), or may even have a reverse effect on pretested subjects (Figure 8.2C). Finally, when there is no interaction, the lines connecting the mean scores for pretested and non-pretested subjects will be parallel, as shown in Figure 8.2D. (Box 8.1 describes a factorial experiment that produced an interaction effect.)
BOX 8.1 Example of a Factorial Design: A Field Experiment on Labor Market Discrimination

Audit studies use experimental methods to investigate discrimination in real-life settings. In a typical study, pairs of auditors or "testers," who differ in race, gender, or age, attempt to rent the same apartments, purchase the same cars, or apply for the same job openings. Although these studies carefully match the pairs on key characteristics such as physical appearance, presentation style, and qualifications, it often is not possible to rule out the possibility that unmeasured observed differences between the testers biased the results. Marianne Bertrand and Sendhil Mullainathan (2004) conducted a field experiment of racial discrimination in the labor market that effectively controlled for such systematic differences between testers.

Rather than use "live" auditors, Bertrand and Mullainathan sent resumes in response to help-wanted ads in Chicago and Boston newspapers. To create realistic templates for the resumes, they altered actual resumes posted on job-search Web sites. Then they employed a 2 × 2 factorial design to investigate the effects of two factors: race (African American or white) and the quality of the resume (high or low). Race was manipulated by varying the name on the resume, with half of the resumes assigned a very white-sounding name (e.g., "Todd Baker" or "Anne Kelly") and the other half a very African American–sounding name (e.g., "Tyrone Jackson" or "Ebony Jones"). Quality of resume was manipulated by systematically varying several criteria such as labor market experience, skills listed, and existence of gaps in employment. In addition, higher-quality resumes were more likely to have certification degrees, foreign language and extra computer skills, and awards or honors.

The experiment took place between July 2001 and May 2002. During that period, resumes were sent in response to over 1,300 "employment ads in the Sunday editions of The Boston Globe and The Chicago Tribune in the sales, administrative support, and clerical and customer services sections" (p. 996). For each ad, two high-quality and two low-quality resumes were sent, with one resume in each pair randomly assigned to a white name and the other to an African American name. The dependent measure was whether an applicant received a phone or e-mail call-back for an interview.

The figure shows both a main effect for race and an interaction effect. First, irrespective of resume quality, African American names were less likely to be called back for an interview than white names. Second, quality of resume made a difference for whites but not for African Americans. Among white applicants, the call-back rate was 11.31 percent with a higher-quality resume and 8.8 percent with a lower-quality resume. Thus, having a higher-quality resume made a statistically significant difference of 2.51 percentage points in whether an applicant received a call-back. Among African Americans, the call-back rate was 6.99 percent with a higher-quality resume and 6.41 percent with a lower-quality resume, which shows a nonsignificant .58 percentage difference. This pattern of discrimination was obtained across occupations and industries in both Boston and Chicago and held as much for employers who listed "Equal Opportunity Employer" in their ad as other employers. The results suggest that racial discrimination still exists and may account for why African Americans continue to do poorly in the labor market.

Besides providing information on interaction effects, factorial designs are cost-efficient. A factorial design can increase the amount of information provided by a study with little increase in cost over a nonfactorial experimental design. Consider the Gottlieb and Carver experiment, for example. By adapting the procedures of the Darley and Latané experiment and manipulating the number of discussants, this study replicated the bystander effect. In effect, it addressed the following question: Does the number of bystanders affect helping in emergencies with a different sample of subjects? But by including another relevant independent variable, anticipated interaction, the factorial design addressed two additional questions: (1) Does anticipated face-to-face interaction affect helping? (2) Will anticipated interaction diminish the impact of the bystander effect—that is, will expected future encounters with bystanders reduce the difference between helping in the absence and presence of bystanders? As it turned out, Gottlieb and Carver did not find an interaction effect, as expected. Rather, there were main effects for both factors: Anticipated interaction reduced the time to respond to the emergency, irrespective of the presence of bystanders, and response time was longer in the bystander than in the nonbystander condition.

A factorial design also may enhance external validity by permitting determination of the effects of a key variable under several conditions. When the effects are consistent under diverse conditions, we are more confident that the findings generalize to additional situations. For example, we may want to study the effects of counseling on troubled marriages, with our dependent measure being the percentage of participant couples still together 1 year after completion of counseling.
A 2 × 3 × 2 factorial design might utilize two counseling approaches let us say "behavior modification" and "eclectic"; three counselor conditions, such as "male-female counselor team," "male counselor only," and "female counselor only"; and two cost-of-counseling conditions, perhaps "fixed fee" and "free." Since this design explores the effects of marriage counseling under twelve conditions, the study should be high in external validity.

Finally, external validity may be enhanced by embedding factorial designs into surveys. Complex factorial designs can be incorporated seamlessly into surveys through computer-assisted interviewing, which we discuss in chapter 9. In this way, the high internal validity of an experiment is paired with the high external validity of a representative sample survey. For example, to explore subtle racial discrimination, Paul Sniderman, Richard Brody, and Philip Tetlock (1991) randomly varied five attributes of telephone interviewers' descriptions of a "laid-off worker" (race, gender, age, marital status, parental status, and work history), after which respondents were asked how much help in finding a new job the government should give the laid-off worker ("none at all," "some," or "a lot"). The 2 × 2 × 3 × 4 × 2 factorial design had ninety-six different versions. In addition to the embedded factorial design, in their analysis Sniderman, Brody, and Tetlock used information about the respondents collected during the telephone survey. One surprising finding was that self-described conservatives were more likely to favor government help if the laid-off worker was black than if she or he was white; the government-assistance position of self-described liberals was unrelated to the racial manipulation.

### Quasi-experimental Designs

Legal, ethical, or practical considerations make it impossible to employ a true experimental design in some research situations. Frequently, random assignment of persons (or other units) is not possible. At other times, control or comparison groups cannot be incorporated into the design. Sometimes, random assignment to treatment and control groups can be carried out but the researcher cannot exercise the tight control over subjects' experiences required for a true experiment. To deal with these problems, researchers have developed a number of quasi-experimental designs, so named because they take an experimental approach without having full experimental control.

In terms of complexity and effectiveness in controlling extraneous threats to validity, these designs generally lie between pre-experimental and true experimental designs. Some of them resemble pre-experimental designs but with added features. Others are similar to true experimental designs but with something lacking, such as a control group or the process of randomization.

An example of the first type is the separate-sample pretest-posttest design (Campbell and Stanley, 1963), diagrammed as follows:

\[
\begin{array}{c|c|c}
R & O_1 & X \\
X & O_2 & \\
\end{array}
\]

\[
\begin{array}{c|c|c}
O_3 & O_4 \\
\end{array}
\]

Except for the missing R representing randomization, this design looks exactly like the pretest-posttest control group design. Randomization, of course, should be used if at all possible. But if it is not possible, then this quasi-experimental design is often worth using. The inclusion of a control group, especially one highly similar to the experimental group in known respects, makes it superior to the one-group pretest-posttest design. For if the groups are similar in recruitment and history, then the design controls for history, maturation, testing, and regression.

Another important type of quasi-experimental design involves time-series data, which consist of multiple observations of the same or similar units over time. The interrupted time-series design resembles the one-group pretest-posttest design except that instead of a single observation before and after the treatment, there are multiple observations before and afterward, as shown below. (Note: There are no limitations, other than practical ones, on the number of observations in the series.)

\[
\begin{array}{c|c|c|c}
O_1 & O_2 & O_3 & X \\
O_4 & O_5 & O_6 & \\
\end{array}
\]
Experimental Designs

The treatment may be systematically introduced into the series by the researcher or it may consist of some naturally occurring intervention. Social researchers often use this design when periodic measurements of an effect of interest are available, such as regularly administered school achievement tests, business records of absenteeism and productivity, or myriad institutional records on events such as automobile accidents, divorces, and crimes. For example, to assess the deterrent effect of a gun-control law mandating 1-year minimum sentences for anyone convicted of carrying a firearm without an appropriate license, one could examine homicide rates severa years before and several years after the law went into effect (see Deutsch and Alba, 1977). If the law had an impact, one would expect an "interruption" or discontinuity in the time series— the homicide rates—at the point where the law was introduced.

The use of several observations makes the interrupted time-series design superior to the one-group pretest-posttest design. Any change from pretest to posttest in the latter design may reflect a long-term trend or a temporary upswing or downswing in a fluctuating pattern. But these effects can be distinguished from treatment effects in a time-series design by comparing the overall pattern with any change at the point of intervention. The biggest threat to internal validity with the interrupted time-series design is history, since it is possible that some more or less simultaneous event other than X may have produced the change. Also, if the time series is taken from official records, an instrumentation effect should be ruled out by checking carefully to see if there has been a change in record-keeping procedures.

The best way to control for history in a time-series analysis is to examine two or more time series, one of which was exposed to the treatment and the others of which were not. This extension of the interrupted time-series design is known as the multiple time-series design.

\[
\begin{align*}
O_1 & & O_2 & & O_3 & & O_4 & & X & & O_5 & & O_6 & & O_7 & & O_8 \\
O_9 & & O_{10} & & O_{11} & & O_{12} & & O_{13} & & O_{14} & & O_{15} & & O_{16} & & O_{17} & & O_{18}
\end{align*}
\]

The subscripts here represent points in time. Now, if a change from \(O_4\) to \(O_5\) occurs in the first group but does not occur in the second group, the change is not likely to be due to a treatment-correlated historical event. This quasi-experimental design is particularly effective in controlling for sources of invalidity. But like the nonequivalent control group design, which is contained within it, the strength of inferences in a multiple time-series design depends on how comparable the control group is to the interrupted group.

Since the design possibilities are so numerous, we limit further discussion of quasi-experimental designs to some general ideas and a detailed description of two quasi-experimental studies that used one or more of the designs outlined above. The objective of such studies, as with true experiments, is to determine treatment effects by eliminating plausible rival explanations of experimental results. Lacking subject randomization or other features of true experiments, however, quasi-experimental studies use a variety of approaches to establish internal validity. Threats to internal validity are ruled out individually in several ways (Cook, Cook, and Mark, 1977): (1) including special design features, (2) examining additional data that bear on each threat, and/or (3) reasoning, on the basis of theory or common sense, that a particular threat is an unlikely alternative explanation.

KEY POINT

Researchers should identify specific threats to the internal validity of quasi-experiments and then consider how these may be ruled out.

An example of a design feature that strengthens causal inferences in a quasi-experiment is the use of a pretest. This is a nonessential feature in most true experiments because randomization creates initially equivalent groups. But in a quasi-experimental study, a pretest permits a vital check on the initial difference between non-equivalent groups (Cook, Cook, and Mark, 1977). An example of the second approach—examining additional data—comes from a study of a smoking-cessation treatment (Berglund et al., 1974). The researchers obtained a measure of motivation from pretreatment interviews with participants. By showing that motivation scores were unrelated to subjects' success in withdrawing from smoking, they eliminated one selection threat as a possible rival explanation. Finally, an example of the third approach would be ruling out testing as a serious threat in an educational setting where the measures are similar to typical classroom testing procedures.

Let us further examine how these approaches are applied in two carefully conducted quasi-experimental studies.

Example 1: Interracial Attitudes and Behavior at a Summer Camp

Can contact between "opposite-race" children under favorable circumstances improve interracial attitudes and behavior? To test this idea, social psychologist Gerald Clore and his colleagues (1978) set up a summer camp for underprivileged black and white children that was designed to provide a positive interracial experience. Previous research suggested that prolonged, intimate contact between racial groups of equal size and similar socioeconomic background could significantly reduce prejudice. Therefore, the camp was structured so that blacks and whites of similar status were equally represented among the campers, counselors, and staff. For example, children were assigned to tents that accommodated three black and three white campers of the same sex and age group (8–10 or 11–12), as well as one black and one white counselor. In all, 196 children were randomly assigned to attend one of five 1-week camp sessions.

The experimenters used three dependent measures to assess whether the camping experience affected racial attitudes and behavior. First, they administered an attitude measure, based on four questions regarding feelings toward children of the "opposite" race. The procedure approximated the separate-sample pretest-posttest design, with half the children administered the measure on the first day of camp and the other half near the end (fifth or sixth day). Results showed the camp to be effective for girls but not for boys. While the girls had somewhat more negative cross-race attitudes at the beginning and shifted toward neutrality, the boys were relatively neutral at the beginning and showed no change.

A second measure, obtained for three of the five weeklong sessions, involved making inexpensive loaded cameras available to the children and allowing them to...
photograph whatever and whomever they wished. Within age, sex, and race categories, children were randomly assigned to either a pretest group that took pictures on the second day or a posttest group that took pictures on the fifth or sixth day. The dependent variable was the proportion of persons in the resultant photographs who were not of the photographer’s race. Thus, this measure, like the first, was based on the separate-sample pretest–posttest design. But unlike the first measure, the photo taking was unobtrusive and behavioral. Analysis of the photos revealed a slight but nonsignificant overall treatment effect: For only one of the three weeks measured did photos taken at the end of the week reveal a significantly higher proportion of cross-race persons than photos taken at the beginning.

A third set of measures, also behavioral, consisted of interpersonal choices the children made in three games played on the first day of camp and again on the fifth or sixth day. Counselors recorded the choices, with the dependent variable being the proportion of choices that were interracial. The design for this measure referred to by the researchers as a “multiple-group pretest–posttest design,” resembled the pre-experimental one-group pretest–posttest design but was an improvement over it in that it was replicated over five camp sessions (“multiple groups”). Results showed a slight shift toward more cross-race choices from the beginning to the end of the sessions. As with the attitude measure, the change was evident for girls only. While showing no significant change from pretest to posttest, the boys made a higher proportion of cross-race choices overall than did the girls.

Thus, the dependent measures, taken together, indicated that the camp experience was effective in changing the interracial attitudes and behavior of the girls but not of the boys, who evidenced more positive attitudes and behavior from the beginning. The authors of the study described how they were able to rule out rival explanations to the hypothesis that the camp experience itself caused the observed changes. To do so, they invoked special design features, additional data, and reasoning based on common sense.

Some crucial design features incorporated in the study enabled Clore and coworkers to eliminate several validity threats. These included the use of two different quasi-experimental designs, three very different dependent measures (attitude questions, photographs, and interpersonal choices), and replications. Using more than one quasi-experimental design strengthens a study because different designs ordinarily will not share the same weaknesses. For example, testing effects cannot be ruled out in the multiple-group pretest–posttest design (which produced the shift in cross-race choices) since pretests and posttests are administered to the same persons. However, testing is not a threat in the separate-sample pretest–posttest design (which produced the change in attitudes) since individuals receive either the pretest or the posttest but not both.

The use of dissimilar measures also strengthens inferences by controlling for systematic error in any one measure. Moreover, the measurement processes were replicated three to five times (weekly camp sessions), controlling for the most part the threats of instrumentation and history. Instrumentation could be ruled out because with pretests and posttests repeated over several weeks, it was unlikely that scorers’ expertise, effort, or enthusiasm would differ systematically from pretest to posttest across all sessions. And since the measurements were replicated at differ-
the crackdown (the treatment) caused the decrease in deaths (the dependent variable), the design would be the one-group pretest-posttest design, which fails to control for most of the common threats to validity.

Because of the weakness of the one-group pretest-posttest design, Campbell and Ross did an analysis based on an interrupted time-series design. Data for this design are graphed in Figure 8.3, which shows the number of traffic fatalities from 1951 through 1959. The decrease in traffic fatalities in 1956 now appears much less impressive in view of the comparable decreases shown in 1952 and 1954 and the unusually high number of fatalities in 1955.

Campbell and Ross then extended their evaluation by using a multiple time-series design. The researchers obtained traffic fatality statistics for the years 1951 through 1959 for four states adjacent to Connecticut, which were assumed to be similar in weather and driving patterns. Figure 8.4 shows the time-series data for the five states. Note that for the year 1955, four of the five states showed an increase in fatalities over the previous year, with Connecticut showing the greatest increase. All five states showed a decrease for 1956. These facts, as well as the fact that the 1956 decrease in Rhode Island closely resembled that in Connecticut, might argue against the hypothesis that the decrease in Connecticut was caused by the crackdown. However,

Connecticut was the only one of the five states to show consistent year-to-year decreases following the crackdown; Rhode Island, by contrast, showed consistent increases after 1956.

Let us now examine how the major threats to validity were dealt with in this study. History is a possible threat in that some event of 1956 other than the crackdown may have caused the reduction in fatalities. Possible rival explanations included better weather conditions in 1956 than in 1955 and improved safety features on 1956-model automobiles (Campbell and Ross, 1968). While the researchers reasoned on the basis of available data that neither of these explanations appeared plausible, the threat of history cannot be ruled out altogether, especially in view of the fatality rate decreases for 1956 in all of the four control states. Some event common to all five states still may have caused the decrease in fatalities. However, history is not a plausible rival explanation for the post-1956 decreases in Connecticut since the adjacent states failed to show the same pattern.

Maturational is normally associated with human physical and psychological processes. In the Connecticut study, maturation would be a threat if, for example, the driving population as a whole were becoming more skilled drivers. Although this is rather implausible, Campbell and Ross extend the idea of maturation to processes external to subjects. Thus, they defined as a maturation threat the possibility of a long-term trend toward a reduction in death rates due to such factors as improved
medical services and improved highways. But since no such trend appears in the time-series or multiple time-series data, this threat may be ruled out.

The threat of testing must also be evaluated. Could the pretest by itself have caused the change? In this case, the pretest consists of the 1955 traffic fatalities statistic. It seems unlikely that mere keeping of records would have much impact on the next year’s statistic, but the widespread publicizing of the high 1955 figure conceivably could have increased driver caution, thus lowering the subsequent year’s statistic. This threat cannot be ruled out with certainty for 1956; however, it seems implausible that publication of the 1955 figure would result in the continued year-to-year decreases that were observed in Connecticut.

Instrumentation would be a threat if there had been a post-crackdown change in record keeping. Campbell and Ross reported that they found no evidence of this.

Regression is a threat whenever a treatment is administered on the basis of a high pretest score. It may be argued that this was the case with the Connecticut crackdown on speeding, which was instituted following an extreme year. On the basis of data from 1951 through 1956, regression or simple instability of the data would indeed offer a plausible rival explanation. However, since the rate of traffic deaths continued to decrease in the years following the crackdown, these factors do not explain the findings adequately. The Campbell and Ross study is an example of rigorous quasi-experimental research. While the absence of a true experimental design made it impossible to exert optimal control over the threats to validity, the researchers gained valuable knowledge by making use of better quasi-experimental designs, relevant available data, and common sense. By these means they were able to conclude with a large degree of confidence that the Connecticut crackdown on speeding had some effect.

Summary

The basic principle of good experimental design is “doing only one thing at a time”—that is, allowing only one independent variable to vary while controlling all other variables. In this chapter we examined pre-experimental designs, true experimental designs, and quasi-experimental designs in light of this basic principle. We found that the pre-experimental designs violate this principle by permitting a number of variables to go uncontrolled, presenting serious threats to the internal validity of the study. Features of the true experimental designs, on the other hand, permit researchers to rule out these threats as rival explanations to the hypothesis. Although quasi-experimental designs control extraneous variables imperfectly, rival explanations frequently may be ruled out through the intelligent use of design features, additional data, and common sense.

Experimental designs are evaluated in terms of how well they control for extraneous variables that threaten a study’s internal and external validity. Thus, we began by identifying several common threats to internal validity: history (specific events other than intended experimental manipulations that occur during the course of an experiment), maturation (psychological and physiological changes in subjects), testing (the effects of being measured once on being measured a second time), instrumentation (unwanted changes in the measuring instrument or procedure), statistical regression (the tendency for extreme scorers on one measurement to move closer to the mean score on a later measurement), selection (differences in the composition of experimental and control groups), attrition (the loss of subjects during the course of an experiment), and interactions with selection (differences between treatment groups due to the combined effect of initial differences and history, maturation, or testing). These threats, or classes of extraneous variables, are sources of invalidity insofar as they can account for study results.

Pre-experimental designs lack one or more features of true experiments, such as a comparison group or random assignment. Therefore, they are subject to several validity threats and their findings cannot be interpreted meaningfully. Three pre-experimental designs are the one-shot case study, the one-group pretest–posttest design, and the static group comparison.

All the basic true experimental designs control adequately for the major sources of internal invalidity. However, the external validity of the pretest–posttest control group design suffers from the possibility of a testing–treatment interaction, in which experimental effects occur only for pretested subjects. This threat is eliminated in the more economical posttest-only control group design, albeit at the loss of posttest information. The Solomon four-group design offers the advantages of both of these designs by combining them in a single experiment. Randomized within-subjects designs offer the advantage of comparing the same individuals in both experimental and control conditions but are seldom used in social research because responses in one condition are likely to affect responses in a subsequent condition. While strong in internal validity, studies incorporating true experimental designs may still be weak in external validity. Besides testing–treatment interaction, external validity may be threatened by the interactions of the treatment with sample characteristics, history, and maturation.

External validity is generally better in factorial designs. These are simple extensions of basic experimental designs in which two or more independent variables are manipulated. Other advantages of factorial designs include their ability to demonstrate both main effects and joint effects and cost efficiency. Each manipulated variable in a factorial design is called a “factor.” Main effects refer to the effects of a single factor by itself. Joint or interaction effects refer to outcomes in which the effect of one independent variable depends on the level or value of another.

Quasi-experimental designs, like pre-experimental designs, lack some feature (usually randomization) of true experiments. Examples are the separate-sample pretest–posttest design, in which both groups receive the treatment but one is tested before the treatment and the other afterward; nonequivalent control group designs, which incorporate control groups but without random assignment; the interrupted time-series design, which involves a series of observations before and after the treatment; and the multiple time-series design, which extends the latter design by adding a series of observations on a nonequivalent control group. By virtue of special design features and supplementary data that test specific validity threats and by rendering threats implausible through reason, quasi-experiments often permit relatively strong inferences about cause and effect. As a result, they often are used to...
assess the effects of social policies and social programs, a subject we will address at length in chapter 14.

**Key Terms**

- pretest–posttest control group design
- posttest-only control group design
- Solomon four-group design
- within-subjects designs
- between-subjects designs
- counterbalancing
- factorial experimental designs
- main effect
- interaction effect
- quasi-experimental designs
- separate-sample pretest–posttest design
- nonequivalent control group design
- interrupted time-series design
- multiple time-series design

**Exercises**

1. Despite the care taken in devising experiments, some manipulations of the independent variable may inadvertently be confounded with other variables. And sometimes this goes unnoticed until after a study is published. See if you can identify the confounded variable in the following study mentioned by Charles Stangor (2004:245). To test the hypothesis that aerobic exercise reduces depression, Lisa McCann and David Holmes (1984) randomly assigned women with mild depression to one of three conditions: an exercise class that met twice a week, a relaxation treatment (placebo) in which subjects practiced muscle relaxation at home, or a no-treatment control group. When all subjects were tested after 5- and 10-week periods, only the exercise group showed a significant reduction in depression. What else, besides exercise, may account for this reduction?

2. For each of the following studies, identify the major threat(s) to internal validity and explain how each threat is an alternative explanation of the study's hypothesized effects.
   a. A campus lecturer touting the benefits of transcendental meditation (TM) presents data showing that persons who have practiced TM for 24 months or longer perform better on recall tests and learn more quickly than a group of randomly selected nonmeditators. According to the lecturer, these data show that TM directly improves the ability to learn.
   b. Stephen Schoenthaler (1983) studied the impact of dietary changes on the antisocial behavior of individuals in juvenile detention facilities. In one study, the diet of inmates was changed by replacing sugar with honey; soft drinks and Kool-Aid with fruit juice; and high-sugar cereals, desserts, and snacks with foods lower in simple sugars. Using reliable institutional incident reports, Schoenthaler measured antisocial behavior for each inmate 3 months prior to, as well as 3 months after, the dietary change. The records showed a significant reduction in behavioral problems.
   c. To test the effects of the children's television series Sesame Street, viewers were tested for various preschool skills (e.g., recognizing letters of the alphabet) both before and after a season of viewing. It was found that children with the lowest initial test scores showed the greatest gains, which suggests that disadvantaged children benefited more than others from watching the program.

3. The authors give an example of an experiment in which the interaction effect of selection and maturation is a threat to internal validity. Imagine the following hypothetical experiment. To study the effects of new writing-intensive courses at a college, a researcher examines two groups of students during a single semester. One group consists of first-year students enrolled in writing-intensive courses, and the other group consists of mostly upperclass students who are not enrolled in such courses. Tests given at the beginning and end of the semester reveal a much greater improvement in writing skills for the group enrolled in writing-intensive courses than for the group not enrolled. How could a selection-maturation effect account for this outcome?

4. One of the nation's most popular programs to control and prevent drug abuse is Drug Abuse Resistance Education (DARE). DARE is a school-based program with a core curriculum that focuses on children in their last year of elementary school (5th or 6th grade), although it has been used with children in kindergarten through 12th grade. Taught by a police officer, the program is designed to increase students' knowledge of drugs, to develop negative attitudes toward drugs, and to enhance students' ability to resist peer pressure to use drugs. Suppose someone tests whether exposure to DARE has these hypothesized effects. Knowing that DARE is being administered in all 6th-grade classes in the local school district, he decides to randomly select ten classes in which to conduct a study. At the beginning of the school year, he measures attitudes toward drug use among all students in the ten randomly chosen classes. Then, at the end of the year, after the children have been exposed to the DARE program, he measures their attitudes toward drug use again.
   a. Use the notation introduced in this chapter to diagram the design of this study. What are the main threats to the internal validity of this study, and how could each threat account for the results?
   b. Suppose the investigator decides to do a second study using an alternative design. He learns about another school district where DARE is not administered and decides to compare the attitudes of 6th-graders in this school district with the attitudes of 6th-graders who are exposed to DARE. In this case, he contacts all the 6th-grade classes from each of the two school districts and measures attitudes toward drug use among all students in these classes at the end of the school year. Diagram the design of this study. What new threat to internal validity, not present in the original study, could account for the results?
c. To carry out an internally valid experimental test of DARE, what features of the study would you need to have control over? Explain.

5. Suppose in the audit study reported in Box 8.1 that the researchers found no interaction effect but that both race and resume quality had statistically significant effects on call-backs for interviews. Graph this outcome in a figure like the one in Box 8.1.

Notes

1. This section and the next three sections draw heavily on Donald Campbell and Julian Stanley's classic treatment of these subjects (1963).

2. Campbell and Stanley (1963) originally used the term "mortality" to describe this source of invalidity. However, "attrition," which is now commonly used, seems better to capture the intended meaning.

3. However, if the teacher initiated the experiment one day because of a sharp increase in talking out of turn, the incidence of this behavior may regress toward a more typical classroom level on the posttest.

4. However, this design does not by itself control for effects of unintended events that might occur within a treatment group. For a discussion of this problem and suggestions for dealing with it, see Campbell and Stanley (1963:13–14).

5. Notice how the principle of "doing only one thing at a time" applies in interpreting factorial results. The only difference between the first-row and the second-row groups is the factor-A manipulation. The two groups are equivalent on factor B (each row has an equal number of pretested and nonpretested subjects) and should be approximately equivalent on extraneous variables as a result of subject randomization.

6. For a more complete discussion, see Campbell (1969b) and Cook and Campbell (1976, 1979).

Survey Research

Survey research in its many forms has become a very common activity in today's world, and most of us have had some experience with it in one form or another. Perhaps you have been stopped on the street by a news reporter and asked your opinion on some issue of local or national importance. You may have responded to a reader survey found in a popular magazine. Or perhaps you have replied to an Internet questionnaire when registering a newly purchased product. You or someone else in your household very likely responded to the last U.S. Census, which attempted to enumerate and gather confidential information about every person living in the United States. You have been a consumer of survey research if you have read in the newspaper the results of Gallup or Roper public-opinion surveys. You have done a little "survey research" of your own if you moved to a new community and asked a number of residents about local restaurants or where to obtain various services.

General Features of Survey Research

These survey examples differ in their degree of formality and in the extent to which they conform to these typical features of professional survey research:

1. A large number of respondents are chosen through probability sampling procedures to represent the population of interest.
2. Systematic questionnaire or interview procedures are used to ask prescribed questions of respondents and record their answers.
3. Answers are numerically coded and analyzed.

As we now elaborate on these three features through reference to actual studies, we will also describe exceptions to the general rule for each feature.

Large-Scale Probability Sampling

Professional surveys make use of large samples chosen through scientific sampling procedures to ensure precise estimates of population characteristics. National opinion polls typically number around 1000 respondents, but surveys of national samples can be much larger. In the National Educational Longitudinal Study (NELS), for example, the probability sample interviewed in 1988 consisted of 24,599 8th-graders from 1052 public and private schools. Participating students were asked