That [continuity and progress] have been tied to careful experimental and theoretical work indicates that there is validity in a method which at times feels unproductive or disorganized.

—Aronson (1980, p. 21)

OVERVIEW

The purpose of this chapter is to provide you with the information you need to evaluate experimental research, specifically, research designed to test cause–effect hypotheses. You will learn about a variety of issues that must be considered when consuming the results of an experiment. For each of the major designs discussed, appropriate questions are suggested so that you can critically evaluate them. General considerations about reading reports are addressed.

INTRODUCTION

To some people, experimental research is the highest peak of scientific research. To others, it is the valley of darkness through which promising scientists must walk before they can do meaningful research. To most researchers, experimental research is the general label applied to methods developed for the specific purpose of testing causal relationships. Other labels include randomized controlled trial, randomized clinical trial, controlled study, and similar phrases that include the words random or control or both. Like Aronson, I sometimes feel that experimental research can be unproductive and disorganized and, at other times, I feel that experimental research includes the best
possible designs for almost anything; experiments, for example, are often called the gold standard in research (Versi, 1992). I never feel as though it is the valley of darkness, but whatever negative feelings I may sometimes have are more than offset by the thrill of finding out why something occurs the way it does. Experimental research may involve the most complicated research designs—that is, until one becomes accustomed to reading it—but it is the only way to obtain a definite answer to the question of why something happens. That is because experimental research is the only way to test causal hypotheses directly. Even though the word experiment is used in a variety of ways in everyday language—it is often used to refer to any kind of research or test—an experiment has some very specific characteristics and the word has a much narrower meaning when used by researchers. The specific meaning when used in the context of research has to do with a process called causal analysis.

**CAUSAL ANALYSIS IN EXPERIMENTAL RESEARCH**

Causal analysis—the logical process through which we attempt to explain why an event occurs—should not be new to you. It is, for example, the basis for explanatory research (see Chapter 1). Within the framework of experimental research, causal analysis includes a combination of three elements—temporal priority, control over variables, and random assignment—the presence of which enables researchers to test cause–effect hypotheses and interpret the results in terms of why something has occurred.

Temporal Priority

One of the requirements of causal analysis is knowledge that the suspected cause precedes the effect. Even though the simplicity of this requirement is readily apparent—something that will happen tomorrow cannot cause something that happens today—the concept can sometimes get a little confusing. For example, the unemployment figures that will be released tomorrow cannot affect today’s decision to invest in the stock market; on the other hand, speculation about what tomorrow’s unemployment figures might be can affect that decision. It is not tomorrow’s event that affects today’s behavior but today’s speculation about tomorrow that affects today’s behavior. Temporal priority, the requirement that causes precede their effects, is a stringent requirement, and we must be careful to understand exactly what is being considered a cause. Figure 6.1 illustrates temporal priority.

Because the requirement of temporal priority is obvious, it is often assumed that temporal priority exists when, in fact, it may not. Consider, for example, the temporal priority involved in Jacobs’s (1967) research on suicide notes discussed in Chapter 1. Jacobs’s content analysis of suicide notes led him to conclude that people committed suicide because they believed the uncertainty of what might happen after death was preferable to the perception of certain, continued depression in their lives. One question Jacobs was not able
to address directly was, “Which came first?” Did people decide to commit suicide because they preferred the uncertainty of death, or did they decide to commit suicide and then justify that decision by writing notes about the uncertainty of death? There is, of course, no way to answer this question using Jacobs’s data; there may be no ethical way to answer this question with any data. Thus, as you read through reports of experiments, look for explicit justification of temporal priority. Merely assuming temporal priority does not count as critical evaluation of experimental research.

Control over Variables

Because temporal priority is often difficult to establish through logic alone, experimental research invariably involves exerting some control over the research environment. Some of that control involves keeping certain things constant, such as the form used to collect the data or the setting (whether inside or outside of a laboratory). Some things cannot be held constant, and they are called, sensibly, variables. One way to establish temporal priority is to manipulate the independent variable—the suspected cause under consideration in a research project. In order to test Jacobs’s hypothesis experimentally, then, we would have to be able to depress a group of people to the point at which they were suicidal and then compare them to a group of people who were not depressed. Obviously, such research would violate just about every known principle of ethics. Let’s continue this discussion with a more feasible experiment.

Oermann, Kardong-Edgren, and Odom-Maryon (2011) exerted control over the amount of practice in their study of cardiopulmonary resuscitation (CPR) skills by assigning some nursing students to a practice condition and other students to a no-practice condition. They ensured that the students in the practice condition received practice by having them go to a skills laboratory each month and engage in six minutes of practice on a mannequin that provided automated advice about the quality of the practice. Students in the no-practice condition were not given opportunities to practice on the voice-advisor mannequin. Thus, their independent variable had two levels: (1) practice and (2) no practice.

The dependent variables, the effects under investigation in the experiment, in Oermann et al. (2011) were measured during a three-minute performance of CPR on a mannequin that could record depth of compression, ventilation volume, and other variables of interest. Because Oermann et al. (2011) had control over the timing of the
independent and dependent variables—they scheduled the testing session several months after the practice session began—they were able to both establish temporal priority and demonstrate control over the variables. If you read Oermann et al. (2011), however, you will not find the phrase *We established temporal priority by . . .*; critical reading requires that we take in the information that is presented and use it to establish temporal priority and control. We do read, for example, that “site coordinators ensured that the study protocol was followed” (Oermann et al., 2011, p. 2), which is how they indicated that they exerted control over the variables, how they made sure that the practice group received practice and the no-practice group did not receive practice. The research hypothesis is illustrated in Figure 6.2.

**Random Assignment**

Despite the use of monitors to control the practice sessions by Oermann et al. (2011), there remain other, plausible explanations for the different CPR skills exhibited by students who received practice and those who did not. It is possible, for example, that students who received practice already had better CPR skills or, perhaps, had even completed a previous CPR course. To attempt to control all of these other possible causes by manipulating them and including them as additional independent variables would soon require more groups of people than would be possible. Instead of attempting to control all other possible explanations through manipulation, investigators rely on **random assignment**, which includes *any procedure that provides all participants an equal opportunity to experience any given level of the independent variable*. In Oermann et al. (2011, p. 2), for example, we read that students were “randomly assigned” to receive either practice or no practice in CPR skills.

Random assignment is a critical part of experimental design because it ensures that participant differences that are not otherwise controlled are equalized across the levels of the independent variable. If there happened to have been some nursing students who already knew how to do CPR, for example, then they would be just as likely to be assigned to the practice group as to the no-practice group. Thus, any differences in the mean skill performance exhibited by the two groups would not be attributable to students who already knew how to do CPR. Similarly, differences between the two groups could not

---

**Figure 6.2** An Example of an Experimental Research Hypothesis in Oermann et al. (2011)

<table>
<thead>
<tr>
<th>Practice (Yes or No)</th>
<th>→</th>
<th>Depth of Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Variable</td>
<td>→</td>
<td>Dependent Variable</td>
</tr>
<tr>
<td>Suspected Cause</td>
<td>→</td>
<td>Effect under Investigation</td>
</tr>
</tbody>
</table>
be attributed to differences in age, proportion of men and women, motivation, intelligence, health, nutrition, strength, or any of the hundreds of other possible reasons why someone might perform CPR better than someone else. Because of random assignment, the only thing that systematically differed among the groups was the amount of practice they received.

**DEMONSTRATION VERSUS DEMOGRAPHY**

The combination of temporal priority, control of variables through manipulation, and random assignment is what makes a research study an experiment, what makes it possible to test cause–effect hypotheses. That same combination, however, tends to produce a somewhat artificial environment. If other institutions chose to use brief practice sessions to maintain CPR skills, for example, there probably would not be monitors in the practice rooms at all times to make sure that the people practicing were really practicing. Other practice rooms might not contain exactly the same make and model of practice mannequins as those used by Oermann et al. (2011). Would those differences necessarily mean that brief practice would not be effective in a different environment? The answer is a resounding no.

Experimental research is not supposed to produce an exact replica of natural phenomena. That’s not heresy but rather a recognition that experimental research has a very specific purpose—to test cause–effect hypotheses—and conclusions drawn from experimental research are drawn about the cause–effect relationship. In Oermann et al. (2011, p. 6), for example, among the conclusions was the line “the findings of this not study only confirmed the importance of practicing CPR psychomotor skills to retain them but also revealed that short monthly practices could improve skills over baseline.” Note the use of “could” in the second half of the sentence. They are not promising that brief practice will improve skills—there is no guarantee. Instead, they are reporting that brief practice has been shown to improve skills.

The issue here is the difference between demonstration and demography. In their experiment, Oermann et al. (2011) demonstrated brief practice improved CPR skills among their participants under the described experimental conditions. Demography, on the other hand, involves the question of how often we can expect those same differences to occur with different participants and under different conditions. Demonstration relies on the extent to which the independent variable is the only systematic difference among the groups. If the mean skills levels are different and the brief practice is the only variable that could have caused those differences, then they have demonstrated a cause–effect relationship between brief practice and skill improvement.

Demography, on the other hand, relies on mundane realism, which refers to the extent to which the experience of research participants is similar to the experiences of everyday life. If others are highly similar to the students who participated in the
research and experienced the same type and amount of practice, then we would expect those individuals also to exhibit improved CPR skills (but there is still no guarantee that doing the same thing would produce exactly the same results). Different people, difference types and amounts of practice, different practice conditions, and other differences might produce different amounts of skill improvement. How much the skill improvement might change would depend upon those other differences and additional research would probably tell us more about that, but for now, we know that brief practice can produce skill improvement. We cannot, however, claim that brief practice will always produce skill improvement. Of course, replicating the experiment—for example, with a different amount of practice or with different students—and obtaining the same set of results would add to the generalizability of the cause–effect relationship we demonstrated in the first experiment. Enough replications would eventually enable us to develop some idea as to just how much brief practice is enough or just what types of mannequins are effective or how long the brief practice needs to be continued. The results of the additional studies would help us understand the demography of brief practice.

**ALTERNATIVE EXPLANATIONS FOR RESEARCH RESULTS**

When the purpose of research is explanation—testing cause–effect hypotheses—every effort must be made to ensure that the independent variable is the only systematic influence on the dependent variable. The results of experimental research typically involve detecting differences among groups as measured by the dependent variable. Therefore, we need to be sure that the independent variable is the only preexisting difference among those groups. Temporal priority, manipulation of variables, and random assignment are the general requirements of an experimental design, but sometimes those requirements are missing from a study that we want to use to draw conclusions about cause–effect. In those cases, we need to consider alternative explanations. An alternative explanation, simply, is any number of defined reason why an independent variable is not the only suspected cause of changes in a dependent variable. As a critical consumer of research, you need to understand alternative explanations before you can determine whether or not a causal conclusion expressed in a research report is warranted.

For the following discussion, let us pretend that Oermann et al. (2011) did not conduct an experiment. Instead, let is imagine they conducted a simple demonstration study in which they asked a single group of nursing students to engage in brief practice of CPR skills and measured their skills before and after practice. For reference purposes, let’s call this the Fake Practice Study. Let’s also assume that the results of the Fake Practice Study were such that CPR skills were better after brief practice than they were before the brief practice. Finally, I should also note that Campbell and Stanley (1963) literally wrote the book on alternative explanations, and much of the following discussion relies heavily on their classic volume.
History Effects

A history effect is a potential alternative explanation to the results of the experiment, produced whenever some uncontrolled event alters participants’ responses. Usually the event occurs between the time the researcher manipulates the independent variable and the time the researcher measures the dependent variable. Sometimes a history effect is caused by a truly historical event, but more often than not, it is produced by more commonplace events.

Suppose, for example, that the American Heart Association issued new guidelines about CPR during the Fake Practice Study. Those new guidelines might have motivated the participants to work on their CPR skills, and it might be that work, rather than the brief practice, that caused the improvement in the students’ CPR skills. Of course, a history effect is not always caused by an historical event. The participants in the Fake Practice Study could just as easily have been motivated by one of their instructors saying, “You know, CPR skills are an important part of nursing,” which is not a particularly historic event. The key aspect of a history effect is that something other than the independent variable happened and that “something” caused the observed effect in the dependent variable.

Generally, random assignment and a control condition enable researchers to eliminate the influence of a history effect. Oermann et al. (2011) had two groups of students, those who did and those who did not receive brief practice, and randomly assigned students to those groups. If there were something that happened—a historical event, an instructor’s comment, or anything else—then it would be equally likely to affect both groups of students. If both groups of students were equally affected, then there would be no group differences in the dependent variable. Thus, random assignment and a control condition don’t eliminate a history effect (that would be impossible), but they do ensure that whatever history effect might have occurred is not the reason for group differences in an experiment.

Maturation Effects

In some sense, maturation is a catchall alternative explanation. Maturation refers to any process that involves systematic change over time, regardless of specific events. From a causal point of view, the passage of time is not the cause of maturation but is merely the most convenient indicator of whatever process may be affecting participants. Most experiments do not last long enough for maturation to occur in the everyday sense of the word—people growing older—but maturation also includes such things as fatigue, boredom, thirst, hunger, and frustration as well as positive outcomes such as greater knowledge, wisdom, enhanced decision skills, and so on. If, for example, in the Fake Practice Study, the yearlong duration of the study convinced the students that CPR skills were very important, then they might start practicing CPR more diligently. Or, if the students became stronger as a result of practice, then they could achieve greater compression depth when tested. Either realization of CPR importance or greater strength would be
a maturation effect in the Fake Practice Study and could account for better CPR skills instead of the brief-practice sessions.

Any design involving random assignment to different conditions provides some protection against maturation effects because all groups should experience about the same amount of maturation, but random assignment alone is usually not sufficient protection. Maturation effects could remain an alternative explanation of results if Oermann et al. (2011) did not ensure that both groups were otherwise treated equally during the experiment, such as testing CPR skills the same number of times in the practice and the no-practice groups.

At this point, you should realize that control over much more than the independent variable is necessary for good experimental research. Not only did Oermann et al. (2011) need to control the independent variable (amount of brief practice), they needed to control the number of testing sessions, ensure that the students’ classes were continued as usual, and make sure that the same mannequins were used. The need for even more control will become apparent as we continue to consider additional alternative explanations of research results.

Testing Effects

You may recall from Chapter 2 that measurement always involves some sort of error; measurement is never perfect. How one phrases questions, for example, can affect the responses one receives. Simply asking a question, the measurement itself, can affect more than the response to the question. In experimental research, testing effects are changes in responses caused by measuring the dependent variable. Testing effects can occur in a variety of ways. One might, for example, measure the dependent variable more than once, thereby creating the possibility that responses on the second measurement reflect memory of the first responses. In the Fake Practice Study, for example, knowing that they are being tested may cause some students to perform CPR better than they would perform if no one was watching. Similarly, testing effects can occur when there is more than one dependent variable: One cannot measure all of the dependent variables simultaneously—one of them has to be measured first—and participants’ responses to the first dependent variable might alter their responses to measures of other variables.

The most obvious means for eliminating testing effects are to measure dependent variables only once and to measure the primary dependent variable before any other measures. Oermann et al. (2011) used mannequins that automatically measured compression depth and rate as well as ventilation volume and rate at the same time. From the participant’s viewpoint, the dependent variable was simply “performing CPR.” Oermann et al. (2011) did measure the dependent variable more than once, but they used different, randomly selected subgroups of students to measure CPR skills at three, six, nine, and 12 months into the experiment. Thus, no students received extra practice at CPR from being measured. When reading research in which the dependent variable
was measured more than one time, read carefully to find out how the investigators dealt with the possibility of testing effects.

**Instrumentation Effects**

Beginning researchers, and even some experienced ones, can become confused about the difference between testing effects and instrumentation effects. Such confusion likely occurs because the two terms seem to refer to the same problem. They are not the same, however, and should be considered separately. **Instrumentation effects** are changes in the manner in which the dependent variable is measured; they are problems caused by inconsistent operationalization of the dependent variable or by inconsistently measuring participants’ responses. Testing effects, on the other hand, are produced by the act of measuring something, even if the measurement itself is consistent.

In the context of the current example, instrumentation effects would be a viable alternative explanation if Oermann et al. (2011) had used different types of mannequins to measure CPR skills. (They did not; they used the same exact make and model mannequin for all measurements.) Different types of mannequins might be more or less accurate, or more or less reliable, and therefore could bias the results. In other experiments, instrumentation effects might be caused by changing the questions on a questionnaire, having different groups use different questionnaires, or not calibrating equipment before each use.

To avoid instrumentation effects, control over operationalization of the dependent variable is critical. The logic of experiments may fall apart completely if those who experience different levels of the independent variable also experience different dependent variables. It may seem obviously foolish to use different versions of the dependent variable for different groups, but there are circumstances that might make such foolishness relatively easy to overlook. Even something as apparently innocuous as differences in the quality of copies of the form used to record the dependent variable can cause instrumentation problems. If one group has copies that are more difficult to read than the other group’s copies, that discrepancy violates the logic involved in having the independent variable as the only systematic difference between the groups. Of course, as a consumer of research, you probably are not going to have access to the kind of detailed information that enables you to determine whether or not instrumentation effects have occurred in any particular study. Nevertheless, you should be attuned to potential clues of instrumentation effects when reading the method section of an article. If researchers use alternative versions of the same scale, perhaps one version before the manipulation and a different version after, then you should make sure the article contains information about the equivalence of reliability and validity of the different versions.

More often than not, instrumentation effects become a problem when the operational definition of the dependent variable depends on someone’s judgment. Subjective measures, such as someone’s rating of the quality of an essay, are subject to various problems. For example, the person making the judgments may grow tired, bored, or careless, and such
changes are, in fact, changes in the dependent measure. In this case, because the instrument is the person making the rating, changes in the rater become instrumentation effects. 

Not allowing raters, judges, or others in a study to be aware of the level of the independent variable experienced by the participant reduces the likelihood of instrumentation effects. A blind rater is someone who is unaware of either the research hypothesis or the experimental group from which the responses came. When you read research reports containing such phrases as the observers were blind to conditions or blind raters were used, it doesn’t mean the observers had a vision deficit. Rather, it means the observers or raters did not know to which experimental group those being observed belonged. In Oermann et al. (2011), the raters were blind because the ‘raters’ were mannequins, so they lacked awareness.

Statistical Regression Effects

In the context of alternative explanations of research results, statistical regression effect does not refer to a particular type of data analysis. Originally described by Galton (1886), statistical regression effect is an artifact of measurement that occurs when extreme scores are obtained and the participant is tested again at a later time. Someone who scores extremely high or extremely low on a measure, if tested again, is likely to obtain a second score that is closer to the average than was the first score. The person’s score is said to regress toward the mean because the score moves back to the average score, either from an extreme high or an extreme low. A less-extreme second score doesn’t always happen; it depends on the amount of measurement error.

Most continuous variables, such as a rating from 1 to 10, an intelligence quotient (IQ) score, and crime rates, include the assumption that the overall distribution of responses should conform to the normal distribution, the bell-shaped curve. Most scores bunch together near the mean of the distribution, and the frequency of scores decreases as the scores become more distant from the mean. Therefore, the probability of obtaining extreme scores is lower than the probability of obtaining scores closer to the mean. Think of the distribution of grade point averages (GPAs) of undergraduates at your school. Most undergraduate students have a GPA somewhere between a 2.5 and a 3.5; the number of students with a 4.0 or a 0.5 is relatively low. Thus, the probability of your running into someone with a 4.0 GPA is considerably lower than the probability of your encountering someone with a 2.5 GPA. If you know someone who earned a 3.9 GPA last semester, statistical regression means that he or she would be less likely to earn a 3.9 or higher this semester, too.

Statistical regression effects, like testing effects, are a problem only when the dependent variable is measured more than once. Only random assignment can be used to avoid them. However, statistical regression effects can become a problem even when an implicit measure of the dependent variable is used. For example, a teacher might select students he or she believes to be the brightest students and give them special assignments designed
to further improve their abilities. If, subsequently, the students exhibit no change in their abilities, perhaps as measured by an alternate-form examination, then the lack of difference could be due either to the fact that the assignments were ineffective or due to statistical regression. That is, some of the brightest, as measured by the teacher’s perceptions, may not be as bright as the teacher perceived them to be (there is measurement error in the teacher’s perception). Subsequent measurement would produce a score closer to average. But if the special assignments are actually making the students brighter, then the combination of independent variable plus regression might produce a net result of no change. Statistical regression brought the scores down, and the assignments brought them back up again, leaving the scores right where they started. Of course, the students did receive some benefit—improved abilities—from the assignments, but statistical regression prevented that benefit from being reflected in their scores. Note that such an experiment would also be subject to instrumentation effects; a teacher’s perception and a written examination are not the same operational definition of the dependent variable.

In general, we should be suspicious about any results in which very high scores became lower, or very low scores became higher, unless the researchers provide information that enables us to rule out regression toward the mean as an alternative explanation. As noted earlier, Oermann et al. (2011) avoided statistical regression effects by selecting different groups of students to receive the repeated measures in their study and by repeating the measurements more than once. Over time, the students in the brief-practice condition improved, and their scores showed continual improvement instead of a one-time change.

Selection Effects

The brightest-student example in the preceding section also involves a specific example of alternative explanations known as selection effects. A selection effect is produced by the manner in which the participants were recruited or recruited themselves. That is, selection effects occur because some characteristic of the participants differs systematically across the groups. In the Fake Practice Study, students who volunteer to participate in the study may also be the kind of students most interested in improving their CPR skills. Once again, random assignment to conditions eliminates selection effects because the selection characteristic, whatever it is, should be equally distributed across the randomly assigned groups. Oermann et al. (2011) included students of different ages and genders, but those differences were overcome by random assignment to conditions. By now, you should be getting the idea that random assignment to conditions is an integral part of any experiment.

Sometimes, however, random assignment is practically or ethically impossible. In such cases, you may read about a technique called matching, which is a second-best alternative. **Matching** is an attempt to equalize scores across groups on any relevant variable. Suppose, for example, Oermann et al. (2011) were not able to assign students randomly to the practice and no-practice conditions in their study, perhaps because
students were free to choose whether or not they wanted to practice CPR skills. Under such conditions, the researchers would probably make sure that they included about the same proportion of men and women in each condition, the same range of ages in each condition, the same distribution of experience in school in each condition, and so on. When there are many different variables on which to match participants, you may read about the use of propensity scores (see, for example, Rubin, 2001), which involves statistical matching. **Propensity scoring** involves statistically matching participants on a variety of variables or measures to rule out those variables as alternative explanations. Although more complicated than simple matching technique, propensity scoring is still a form of matching.

The issue with using matching to overcome selection effects is, very simply, that we can never be sure that the researchers have included all of the relevant variables in the matching process. If someone were to replicate Oermann et al. (2011) without random assignment, they would have to match the students in terms of gender, age, amount of previous education, amount of experience with CPR, arm strength, general physical condition, and so on. There may be no limit to the number of different things that might affect CPR skills, so there may be no limit to the number of different variables on which the participants would have to be matched. Even though matching does provide an alternative when random assignment is not possible, it is not as effective as random assignment. Thus, we should be much more critical, and skeptical, about tests of cause–effect explanations that include matching instead of random assignment. Sometimes matching is the only available technique to a researcher, but that doesn’t mean that it is as good as random assignment in controlling selection (or other) effects.

**Attrition Effects**

Attrition effects are changes in the dependent variable caused by the loss of participants during a research project. (Sometimes the loss is caused by the death of participants during the project, in which case the term mortality effects is used.) Attrition is a specific type of selection effect, one due to participants’ leaving, rather than joining, the research project. Oermann et al. (2011) did have attrition in their study; some students quit participating over the 12-month course of the experiment. That attrition occurs in a study, however, does not mean that attrition effects have occurred. That is, just because some participants left the study doesn’t mean that the study is ruined, but it does mean that we should read critically to find out what the researchers did to rule out attrition effects.

Random assignment is considered a safeguard against attrition effects because the number of participants likely to drop out of a study should be roughly equal across groups, as was the case in Oermann et al. (2011). Random assignment, however, cannot be considered a cure for attrition effects, for one experimental group could in fact contain a disproportionate number of dropouts. When you read a study with disproportionate
attrition among conditions, then you should add the phrase *for those who remained in the study* to any interpretation of the results. As consumers, we cannot know the reasons for the disproportionate attrition, but we can wonder about attrition effects as an alternative explanation.

When you read a study in which attrition effects are likely to be of concern, either because the procedure was long or there are multiple sessions, read carefully to find mention of comparing the dropouts across conditions or comparing the dropouts to the remaining participants. Look for results in which the researchers compared proportions of genders, ages, or other variables. Look for wording such as *the average age of the attrition group was the same as the average in the continuing group or those who withdrew consent in the experimental group were slightly older than those in the control group.* Such comparisons will enable you to decide whether or not attrition effects are a reasonable alternative explanation for the results.

Participant Bias

We have already noted that research participants can be affected by *self-presentation*, the concern for the impression one makes upon others. This is one example of *participant bias*—any intentional effort on the part of participants to alter their responses for the purpose of self-presentation. Another example may involve participants’ concerns about revealing sensitive information simply because they believe it is none of the researcher’s business what they think. Generally, any *evaluation apprehension*—concern about being observed—can produce participant bias. In the Fake Practice Study, participants might be concerned that poor CPR performance could affect their grades in their courses and be more motivated to perform well. Thus, the results would be due to participant bias instead of brief practice.

Random assignment to conditions helps to reduce participant bias because, as usual, random assignment equalizes the distribution of apprehensive participants across experimental groups. The only way to avoid participant bias completely is to prevent the participants from being aware that they are being observed, but this invokes the ethical issue of disregarding informed consent. Participants sometimes are unaware that they are in a research project, but this occurs very infrequently.

In addition to evaluation apprehension and related forms of participant bias, participants may intentionally attempt to help or hinder the research efforts. The *beneficent subject effect* occurs when participants are aware of the research hypothesis and attempt to respond so as to support it. Suppose, for example, that students in the Fake Practice Study wanted to perform CPR well because the researcher was one of their favorite faculty members. In this case, the results would be caused by participant bias instead of brief practice. The opposite effect, known as the *maleficent subject effect*, occurs when participants are aware of the research hypothesis and attempt to respond so as to undermine it.
The only way to prevent either effect is to prevent participants from becoming aware of the research hypothesis. This is known as keeping the participants “blind” and is analogous to keeping raters of subjective dependent measures in the dark about the hypothesis. In some projects, a double-blind procedure is used; the raters (or investigators) and participants are unaware of the research hypothesis or of group memberships related to that hypothesis. Blind and double-blind studies may involve concealment of the research hypothesis by preventing awareness of participation itself or they may involve some sort of deception (providing false information about the research project). It is common to keep participants and raters unaware of the specific hypotheses in research, but deception poses ethical problems and is considered only after it has been determined that simply withholding information about the specific research hypothesis will not prevent participant bias.

**Experimenter Bias**

Participants are not the only people who may alter their behavior during an experiment. Experimenter bias refers to the researcher’s differential treatment of experimental groups. There was, for example, considerable potential for experimenter bias in Oermann et al. (2011). The researchers and site coordinators could have especially encouraged students in the brief practice groups to improve their CPR skills. Oermann et al. (2011) conducted their study across ten different schools, however, and the chances that such a breach of research protocol happened at all ten schools are extremely low.

The logic of the experiment requires researchers to treat all groups exactly the same; the only systematic difference between the groups should be the manipulated independent variable. As you read, therefore, pay particular attention to efforts to avoid experimenter bias and all of the previously discussed alternative explanations. Just because researchers are not blind to conditions, for example, does not mean that there was experimenter bias exerting effects upon the results. Experimenters who are not blind, participants who are not randomly assigned, or any of the other aspects of the research that might alert you to potential alternative explanations, however, should be considered very carefully so that you, as the consumer of the research, can decide whether or not any of these alternative explanations pose a problem for interpreting the results. You should not assume there is experimenter bias simply because the results section did not contain the phrase the experimenter was blind to conditions, but you also should not assume that there are no alternative explanations just because the study was published.

**EXPERIMENTAL DESIGN**

Fortunately, not every experiment is subject to every alternative explanation described above. On the other hand, every alternative explanation must be considered a potential problem until logic, control, or experimental design enables you to rule it out.
In this section, we’ll consider the various experimental designs that can be used to rule out alternative explanations. **Design** refers to the number and arrangement of independent variable levels in a research project. Although all experimental designs involve manipulated independent variables and random assignment, different designs are more or less efficient for dealing with specific alternative explanations.

The design a researcher uses depends upon the research hypothesis the researcher has tried to test. Therefore, being familiar with a variety of different designs enables you to critically consume a variety of research projects as you attempt to inform your practice empirically. The major factor in examining a design critically is not its complexity but the extent to which it provides internal validity. **Internal validity** refers to the extent to which the independent variable is the only systematic difference among experimental groups (Munn, Lockwood, & Moola, 2015; Shadish, Cook, & Campbell, 2002). That is, the internal validity of an experiment allows you, as the research consumer, to conclude that the independent variable produced the effects measured with the dependent variable. Just as every poker hand either wins or loses the pot, every design is either a winner or loser at internal validity, depending on the specific research hypothesis being tested.

### The Basic Design

The basic design is the simplest design that still qualifies as a true experimental design. Campbell and Stanley (1963) refer to it as the **posttest-only control group design**, while others call it a **randomized trial** or a **randomized comparative trial**. If we were to add a no-practice group to the Fake Practice Study and randomly assign volunteers to receive either practice or no practice, then we would have an experiment with the basic design, as illustrated in Figure 6.3. When one of the levels of the independent variable is a **control condition**—the absence of the manipulation—then the design is sometimes called a **randomized control trial**. Sometimes researchers will use a **placebo**, a treatment that appears to be real but is not effective, instead of a control group, and the design may be called a **randomized placebo trial**. For example, instead of “Do Not Engage in Brief Practice” in Figure 6.3, researchers might have participants engage in a brief practice of inserting an IV or some other activity that is not related to CPR. That way, both groups engage in about the same amount of activity, but only one group practices CPR skills.

The basic design is not necessarily limited to two groups. Researchers could include brief practice of CPR, brief practice of IV insertion, and no practice at all in a single study. The key aspects of the basic design are that participants are (1) first randomly assigned to two or more groups and (2) measured on the same dependent variable.

The basic design is most efficient for research in which the pre-manipulation state of participants—what they are like before they experience a single independent variable—is either not of interest or can be assumed to be unrelated to the independent variable. Neither change over time nor differential reactions to the independent variable as a
function of some preexisting characteristic can be studied with the basic design because there is no way to know anything about the participants before they experience the independent variable.

The Basic Pretest Design

The basic pretest design, as the name implies, involves adding a pretest measure to the basic design. A pretest is a measure of the dependent variable that is administered before intervention or treatment so that change can be determined. The obvious reason for using a pretest measure is to examine how much the independent variable causes participants to change. The basic pretest design, illustrated in Figure 6.4, is the design Campbell and Stanley (1963) call the pretest–posttest control group design. In Figure 6.4, our fictitious researchers would measure everyone’s CPR skills, the randomly assign the participants to engage or not engage in brief practice, and then measure the same CPR skills again. Change would be measured by comparing skills displayed after the intervention to the skills displayed during the pretest. Although not depicted in the figure, the basic pretest design can include more than two levels of the same independent variable.

The obvious advantage of the basic pretest design over the basic design is the ability to obtain information about the pre-manipulation state of the participants, to examine the change in scores. The disadvantage is that the pretest measure may affect participants’ reactions to the independent variable—that is, asking participants first to display their CPR skills could sensitize them to take greater advantage of the practice opportunities later in the experiment, which could produce a type of testing effect.

Random assignment enables us to overcome the possibility of general testing effects, but the combination of pretest measures and manipulation of the independent variable may create another alternative explanation for the results. Campbell and Stanley (1963) call this alternative explanation a testing–treatment interaction, in which
participants experiencing one level of the independent variable may be more sensitive to testing effects than participants experiencing a different level of the independent variable. Essentially, the pretest measure may make one of the levels of the independent variable, such as the brief practice, more forceful than it would have been without the pretest. This increased forcefulness of that particular level of the independent variable, then, is an artifact rather than a valid test of the variable’s impact. The dependent variable in such cases measures both the effect of the independent variable and its combination with the pretest instead of measuring the effect of the independent variable only. It is also possible that the pretest measure may make participants wonder about the purpose of the study and increase participant bias.

The Solomon Four-Group Design

The most effective design for dealing with the problem of testing–treatment interaction is the Solomon (1949) four-group design, illustrated in Figure 6.5. Although there are only two levels of the independent variable, four groups are required to assess the extent to which testing effects have occurred. It is important to realize, however, that this design does not eliminate the testing–treatment interaction but rather enables the researcher to determine whether or not it has occurred and, if it has, assess its impact. This design also enables one to determine whether or not overall testing effects have occurred; it enables the researcher to assess the effectiveness of the random assignment procedure.

When there are control conditions, then comparing the control with pretest (Group 2) to the control without pretest (Group 4) provides a test of the overall testing effect.
The only difference between these two groups is the existence of a pretest measure, and any difference between these two groups would be due to the pretest. Comparing the difference between change in Group 1 and change in Group 3 provides an estimate of the testing–treatment interaction effect. If there is no interaction between testing and treatment, then the effect of the treatment should be the same with and without a pretest. If, however, there is a testing–treatment interaction, then you will probably read about statistical analyses used to adjust for the interaction. Critical reading, however, will also involve being able to notice when a four-group design should have been used but was not used.

Unfortunately, the Solomon four-group design is not, in general, a very efficient design. It requires twice as many groups as the basic pretest design to examine essentially the same cause–effect hypothesis. The four groups depicted in Figure 6.5, for example, include only two levels of a single independent variable. The loss of efficiency is related to the need to test for the testing–treatment interaction. In general, the more the researcher needs to know, the more groups or participants will be required. Oermann et al. (2011), for example, used a version of the Solomon four-group design by randomly selecting some participants to be measured at various times throughout their yearlong study. Because they had many participants, they were able to check for a testing–treatment interaction effect even though they did not actually include a pretest in their study.

Factorial Design

Many research questions require inclusion of more than one independent variable in the design. Hagger, Wong, and Davey (2015), for example, manipulated both mental simulation (or visualization) and self-control building in their study of mechanisms for reducing binge drinking among college students. Their design allows participants...
Designs that include more than one independent variable are called factorial designs. In terms of our diagram scheme, a simple factorial design is illustrated in Figure 6.6. Hagger et al. (2015) manipulated mental simulation by having participants either (a) visualize behavior related to reducing alcohol consumption or (b) visualizing behavior related to going to the movies or going shopping. They manipulated self-control building by having participants regularly engage in (a) tasks that involve challenging exercises in self-control or (b) tasks that involve very easy exercises in self-control. Within the design, participants collectively experience all possible combinations of the two independent variables, but each participant experiences only one of these combinations. As with other designs, each of the independent variables can have two or more levels.

Often, the notation used in Figure 6.6 is not applied to factorial designs. As you probably discovered while attempting to decipher Figure 6.6, the notation is a little cumbersome. Instead, the notation used in Figure 6.7 is adopted for factorial designs in which the combination of independent variables forms a kind of matrix. I have added identification numbers to the groups, called cells in the design, to make it easier to refer to them in further discussion, but you will usually see means, standard deviations, or some other summary statistics when reading research articles.

In Cell #1, participants will receive irrelevant mental simulation instructions and will be presented with easy self-control tasks, which serves as the control condition. In Cell #2, the mental simulation is irrelevant but the self-control task is challenging; this cell, together with Cell #1, is comparable to a basic design for self-control building alone. In Cell #3, participants engage in relevant visualization and are presented with easy...
Evaluating Research

Figure 6.7  The Factorial Design of Hagger et al. (2015) in Matrix Notation

<table>
<thead>
<tr>
<th>Type of Visualization</th>
<th>Type of Self-Control Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
</tr>
<tr>
<td>Movies or Shopping</td>
<td>1</td>
</tr>
<tr>
<td>Reducing Drinking</td>
<td>3</td>
</tr>
</tbody>
</table>

self-control tasks; this condition, together with Cell #1, is comparable to a basic design for mental simulation alone. Finally, participants in Cell #4 engage in relevant visualization and presented with challenging self-control tasks. Cell #4 enables the researchers to assess the combination of both independent variables. The design depicted in Figure 6.7 is called a 2 × 2 factorial design—two levels of one independent variable combined with two levels of another independent variable.

The advantage of a factorial design is that interactions between independent variables can be tested. An interaction is a result that occurs when the effect of one variable depends upon which level of another variable is present. Interactions are sometimes called moderator effects (Baron & Kenny, 1986). Hagger et al. (2015) noted that a number of researchers have demonstrated the utility of mental simulation on reduction of unwanted behavior, including binge drinking, and a number of other researchers have demonstrated the utility of self-control building on improving healthy behavior, but they reported that no studies have examined the combination of both strategies. That is, they expect that the effect of mental simulation will be even greater when paired with self-control building. It is equally reasonable to write that they expect the effect of self-control building to be even greater when paired with relevant mental simulation.

Figure 6.8 contains four different fictitious results from Hagger et al. (2015). In Panel A, neither type of visualization nor type of self-control task has much of an effect upon the number of binge-drinking episodes per month. In Panel B, it is clear that mental simulation about reducing drinking does reduce the number of binge episodes compared to visualization about shopping. This is known as a main effect for the type of visualization. Generally, a main effect occurs when different levels of one independent variable evidence different results in the dependent variable. Panel C contains fictitious results for a main effect for type of self-control task. Panel D in Figure 6.8 illustrates an interaction effect. There appears to be a main effect for type of visualization and for type of self-control task, but notice that visualizing drinking less combined with a challenging self-control task reduces the number of binge episodes more than one would expect from the mere combination of both main effects. That is, mental simulation and self-control training together produce an extra strong effect, a greater effect than either variable could produce on its own. The nonparallel lines in Panel D are another indicator of an interaction effect.
The ability to detect and interpret interactions is the primary advantage of factorial designs. As you read articles, keep in mind that no matter how complicated a design may appear to be, factorial designs include only two types of effects, main effects and interaction effects. Don’t allow yourself to be overly confused when trying to understand complicated factorial designs; interpret the results one main effect at a time and one interaction at a time.

Within Subjects Design

A within subjects design is a design in which the same participants are exposed to more than one level of an independent variable. This is often done so that each participant (sometimes called subject) serves as his or her own control; that is, the independent
variable is manipulated within, rather than between, the participants. A within subjects design also involves repeated measures design, in which the dependent variable is measured more than once for each participant. The basic pretest design, for example, is the simplest form of repeated measurement design, but that design does not involve exposing any participant to more than one level of an independent variable.

Kim, Gordon, Ferruzzi, and Campbell (2015), for example, were interested in testing whether or not the lipid content of eggs would enhance absorption of carotenoids contained in vegetables. Their participants consumed a salad to which zero, one and a half, or three scrambled eggs were added and provided hourly blood samples for 10 hours after consumption. The participants consumed each type of salad on different days, thus being presented with all three levels of the independent variable. Kim et al. (2015) also employed repeated measures of their dependent variables by drawing blood from the participants over a ten-hour period. Figure 6.9 contains an illustration of the major parts of the design, which included random assignment to the different orders in which the participants consumed the salads.

As is evident from Figure 6.9, each participant consumed all three types of salads, but the salads were consumed in different orders so that Kim et al. (2015) could assess whether or not the order of consumption made a difference. This is known as counterbalancing, a methodological tactic in which all orders of levels of an independent variable are employed in a within-subjects design to assess the impact that order may have on the dependent variable. When reading studies in which within subjects designs are used, always look for an indication that counterbalancing was used; without it, there is always the possibility that the order of the levels of the independent variable had an impact on the dependent variable.

For some variables, repeated measures designs are simply not possible. Some independent variable effects may last so long that they interfere with later, different levels of the same variable. Kim et al. (2015), for example, waited 14 days between consumption of the different salads to ensure that lipid intake from a preceding salad did not affect the absorption of carotenoids. Critical reading of within subjects designs also includes consideration of the amount of elapsed time between different levels of the independent variable.

Participant Characteristics

Before we leave design, let’s consider the use of participant characteristics as independent variables. Participant characteristics, sometimes called subject variables, are variables that differentiate participants but cannot be manipulated and are not subject to random assignment. Participant characteristics include such variables as gender, age, ethnicity, amount of formal education, height, and so forth. They can be included in an experimental design, but because they are not subject to manipulation or random assignment, they cannot be considered true independent variables in an experimental design.
Figure 6.9 Within Subjects Design Used by Kim et al. (2015)
When reading research reports, however, you will often find that they are described as independent variables or sometimes called subject variables. Oermann et al. (2011), for example, included gender and degree type (diploma, associate, baccalaureate) as participant characteristics in their study of brief practice for resuscitation skills among nursing students. They obviously could neither manipulate nor randomly assign students to gender or degree groups, but they did include both variables in their initial analyses to assess the extent to which men and women and different degree aspirations might be related to performance of resuscitation skills (neither variable evidenced a relationship with performance).

When participant characteristics are included in an experimental design, conclusions about cause–effect relationships cannot be drawn from any effects associated with such variables. Despite this restriction, you would not have to look very long before finding a research article in which the author(s) did exactly that. Drawing cause–effect conclusions about participant characteristics seems to be an almost irresistible temptation to many researchers. When this happens, it usually results from a very logical consideration of the effect and the researcher’s knowledge about related research. Suggesting potential explanations for a gender effect, for example, is certainly within the realm of scientific research. On the other hand, concluding that a gender effect results from, say, differential attitudes when attitudes have not been manipulated in the design falls well outside the logic of experimental research.

DEMONSTRATION VERSUS DEMOGRAPHY AGAIN

Earlier in the chapter, the primary purpose of experimental research was described as testing whether or not a cause–effect relationship can be demonstrated. This purpose does not automatically rule out generalizing the results of the experiment, but generalization (demography) is secondary to testing the relationship (demonstration). If generalizing well beyond the experimental environment is an important part of your intentions as a consumer, you need to ensure that your efforts in that direction are not affected by the internal validity of the experimental design. If random assignment is not consistent with generalization, for example, then you should not generalize as though random assignment didn’t occur.

Overgeneralization is also something to avoid. Although overgeneralization is a potential problem in any research method, experimental research seems particularly prone to the phrase research has proved. Random assignment is critical to experimental research, but experimental research is a process that also depends on replication for its effectiveness. Like any procedure based on probability theory, random assignment works in the long run but may not be effective on a one-time-only basis. Any research requires replication before we can rely heavily on the results.

You should realize that a single experiment does not prove that a cause–effect relationship exists; rather, it demonstrates the existence of the relationship under the
conditions created by the experimental procedures. Those conditions include the specific experimenter, participants, operational definitions, and a host of other potential factors that differ from one experiment to another. A demonstration that something can happen does not mean it always will happen. Consuming a valid experiment requires paying attention to all aspects of experimental research, and that includes looking for additional studies in which the cause–effect relationship has been demonstrated under a variety of different circumstances.

SUMMARY

- Experimental research methods are the only methods designed specifically to test cause–effect hypotheses. Experiments are accomplished by manipulating the independent variable, randomly assigning participants to the various levels of the independent variable, controlling or eliminating alternative explanations, and measuring responses via the dependent variable. The independent variable is the suspected cause; the dependent variable is the effect of interest.

- Generalizing the results of an experiment well beyond the experimental situation is logically impossible, for the major purpose of most experiments is to demonstrate that the cause–effect relationship can occur, not that it always occurs.

- The logic of experimental research is that any difference between groups of participants as measured by the dependent variable is caused by their different experiences with the independent variable. Therefore, an experimenter must maintain internal validity—he or she must rule out alternative explanations of any obtained differences.

- Alternative explanations are generally ruled out through the use of random assignment to conditions created by manipulating the independent variable. These alternative explanations include history effects, maturation effects, testing effects, instrumentation effects, statistical regression effects, selection effects, attrition effects, participant bias, and experimenter bias.

- Control over the experimental situation can be used to rule out instrumentation effects, participant bias, and experimenter bias.

- The basic design of an experiment includes different groups representing different levels of a manipulated independent variable to which participants are randomly assigned. Adding a pretest to this design enables us to measure change as a function of the independent variable. Care must be taken, however, to avoid an interaction between treatment and testing.

- The Solomon four-group design can be used to measure testing effects, including a testing–treatment interaction. This added ability to test effects decreases the efficiency of the design with respect to testing the research hypothesis.
• When more than one independent variable is of interest, factorial designs are used to assess both main effects and interaction (moderator) effects. Main effects are simple effects due to one variable, whereas interaction effects are those caused by a combination of two or more independent variables.

• Within subjects design measures may be used with any experimental design, but only if the effects of an independent variable are not so long lasting as to interfere with subsequent levels of the same or another independent variable. Within subjects designs, which almost always include repeated measures of the dependent variable, should also take into account the possibility of order effects.

• Although often used in experimental research, participant characteristics cannot be considered valid independent variables. Also called subject variables, they may indicate the presence of a systematic difference, but they are not themselves considered to be causal agents.

EXERCISES

1. Find a research article in which the authors identify the research as an experiment. Determine whether the investigators established temporal priority and used random assignment.

2. Using the same or a different article, identify the design used in the research.

3. Using the same or a different article, examine the design and procedure carefully for each of the alternative explanations described in the chapter.

4. Find an article in which the authors describe an interaction (moderator) effect. Identify the variables involved in the interaction and try to explain the interaction to someone who has not read the article.