EVALUATING SOCIAL SCIENCE RESEARCH

Second Edition

Paul C. Stern
National Research Council

Linda Kalof
State University of New York, Plattsburgh

Oxford
NIVERSITY PRESS

3

EVALUATING SCIENTIFIC EVIDENCE: WHAT CONCLUSIONS FOLLOW FROM THE EVIDENCE?

You are now ready to begin evaluating scientific evidence. You have learned how to distinguish scientific evidence from nonscientific statements (Chapter 1) and how to discriminate among the major methods of gathering scientific evidence (Chapter 2). You may also have attempted to formulate a question that is answerable by scientific evidence and to find some of the evidence relevant to your question (using the strategy outlined in the Appendix). If so, this evidence may now be before you. The rest of this book is devoted to teaching the skills that will allow you, once you have found the evidence, to decide for yourself what conclusions to draw from it.

To get an idea of the dimensions of this task, consider a piece of research. Glock, Ringer, and Babbie (1967) theorized that the function of churches in our secular society is to compensate people for their social deprivations. This theory suggested the hypothesis that the more socially deprived people are, the more involved they will be in church activities. To test this hypothesis, the authors evaluated social deprivation and church involvement in a sample of Episcopalians from 234 congregations. Social deprivation was measured on a zero to eight point scale with people being given points for being female (two points), over 30 (one point), over 50 (one more point), unmarried (one point), childless (one point), and middle class (one point) or lower class (two points). The total points assigned to a person was the social deprivation score. Church involvement was measured on the basis of subjects' responses to questions about their participation in specific church activities. The researchers found that the higher a person's social deprivation score, the higher the index of church involvement. The results supported the hypothesis, and were taken as evidence for Glock's theory of the function of churches. (Glock and his colleagues presented more evidence than this, but this summary is enough for our purposes.)

The authors' conclusion is clear, but we should not accept these conclusions without first checking them ourselves. It is proper to ask two kinds of questions about a study before we accept its conclusions:

1. Do the data support the authors' conclusions with respect to the population studied? (In this study, we ask whether social deprivation influenced church involvement among U.S. Episcopalians in 1952, the year the data were collected.)

2. If the conclusions are sound, do they generalize beyond the population sampled and the setting studied? (We must ask whether Glock's results generalize to other churches and to other years. Remember, Glock theorized about *all* churches, but studied only one religious denomination.)

The first question involves what Campbell and Stanley (1963) have called *internal validity*. A study has internal validity to the extent that the data support conclusions about the hypothesis in the specific instance studied. We make judgments about internal validity by examining the details of the study itself and the reasoning that the author used to draw conclusions from the evidence. For instance, we examine the procedural details of the study to decide whether the procedures used to measure and manipulate variables faithfully represented those variables. The Glock study has no internal validity unless the social deprivation scores actually measure social deprivation. For example, if the assumption that people over 30 are more socially deprived than people under 30 is false, the operational definition of social deprivation is not measuring social deprivation, and it follows that conclusions about social deprivation cannot be drawn. The same kind of problem exists with the measure of church involvement. When people are asked about their participation in church activities, they sometimes say what they feel they should have participated in, rather than what they actually did. If this happens to any great extent, the "index of involvement" may be more an index of guilt about church activities, and conclusions about church involvement cannot be drawn. These illustrations should suggest that procedural details can make a difference in what a study is actually measuring. It is difficult to judge the internal validity of a piece of research.

A careful look at the conclusions the author drew from the evidence can sometimes reveal errors of reasoning that undermine the conclusions that seem to follow from the study. Suppose that the Glock study does not have measurement problems—that the scales actually measure what the authors claim. If so, the study would show that socially deprived Episcopalians are more strongly involved in church activities than are others of their faith. But it does not show that social deprivation was a *cause* of their church involvement. And even if the Glock study, combined with other evidence, convinced you that social deprivation did cause these people's church involvement, knowing this would not be enough to support Glock's hypothesis, even for Episcopalians, that "the function of churches . . . is to compensate people for their social deprivations." If we somehow knew that social deprivation caused people to become more involved in church, we could only conclude that compensating for deprivation is *one* of the functions that church involvement has for socially deprived people. We would have learned nothing about what other functions church might have for them, or for other people, or about why people who are not socially deprived join churches.

Glock's study illustrates that even when research produces just the results the researcher expected, the study may not mean what it seems. It is particularly easy to jump to the conclusion that if a study is consistent with its hypothesis, the hypothesis must be correct. In reading research reports, look out for such errors in reasoning. It is especially important to be alert when the research results support your own preconceived ideas, because that is when we are least critical of our own reasoning or someone else's.

The purpose of this chapter is to teach you which questions to ask about the procedures and observations used in a study so that you can make your own judgment of a study's internal validity. Although we do not emphasize the kinds of reasoning problems just mentioned, you will learn how to scrutinize research and uncover many kinds of flaws in imperfect research (and if you look closely enough, almost all research is imperfect). By looking carefully at the pitfalls in research you will increase your ability to make independent judgments about the factual claims that researchers make, and also gain an appreciation of well-conducted research when it can be found. In the exercises and problems, we give you summarized research reports to analyze and evaluate in terms of the concepts presented here. (Chapter 4 provides practice in using the same skills to evaluate actual reports from professional journals.)

The second question mentioned above concerns whether the findings of a study can be generalized to other populations and settings. This involves *external validity*, and it only has meaning once the internal validity of a study has been established. A study has external validity to the extent that its results can be generalized to other situations in which the same variables operate. Thus, Glock's findings have external validity if they hold for other churches in other times, and for other types of social deprivation besides those Glock measured. As this example may suggest, external validity is best determined by comparing the findings of different pieces of research about the same variables. The evaluation of external validity is discussed in Chapter 5.

According to the working definition presented above, a study has *internal validity* when it is possible to draw conclusions about the hypothesis from the data. Internal validity depends on the link from concrete observations to the abstractions they are supposed to be related to; from operational definitions to their corresponding variables. For example, Glock's study has no internal validity unless people's *reports* of their church activities reflect their actual involvement in the church. If these reports actually reflect a desire to look like a good Christian, or to please the interviewer, or to conform to behavioral standards in the community, the authors have measured not church involvement, but some *extraneous variable*.

An *extraneous variable* is a variable capable of explaining the findings of a study without invoking the hypothesis. In other words, the presence of an extraneous variable allows for *alternative explanations* of a set of observations: either the observed relationships are due to the variables in the hypothesis, *or* they result at least in part from an extraneous variable.

A researcher's central problem in demonstrating the internal validity of a piece of research is to achieve control over extraneous variables—there must be a way to rule out alternative explanations of the findings. For example, here is an alternative explanation of Glock's results: women in our society are supposed to be a religious influence in the family, so they may claim more church involvement than they actually have. Because the operational definition of deprivation gives women higher scores than men, this explains why people who score high on deprivation also score high on church involvement. This alternative can be ruled out if high social deprivation scores are related to high involvement scores when only women (or only men) are compared. In such groups, the relationship of social deprivation scores to church involvement scores cannot be due to gender differences.

The task of judging internal validity is the task of interpreting evidence. In Chapter 1, it was noted that "for observations to have scientific value, they must reliably concretize abstractions." Thus, useful evidence must be in concrete language. But there can be many ways that evidence may relate to the abstract variables we are really interested in. So the logical jump from concrete evidence to abstract variables is crucial to the scientific process. To evaluate a scientist's report of research, one must identify the scientist's conclusion from the evidence (the hypothesized explanation), and compare it with other possible conclusions (alternative explanations). Internal validity increases as these alternative explanations can be ruled out.

This chapter provides a guide for finding alternative explanations for social scientists' findings. It introduces some common extraneous variables and gives information about where to expect them and how they may be controlled. You will come across many new terms. Keep in mind that your purpose is not to memorize the terms but to get a feeling for the extraneous variables that exist in various types of research, so that you can suggest alternative explanations for research findings you read. Your primary goal is to learn to analyze and evaluate research reports.

This guide is organized around the seven methods of gathering evidence presented in Chapter 2. Because each of these methods has its own procedures, each has its own characteristic extraneous variables. Consequently, your search for alternative explanations of the evidence will take different directions depending on the method used to gather the evidence. Some extraneous variables are almost universal problems in scientific research, while others cause difficulty primarily with particular research methods. The most basic research method (naturalistic observation) tends to raise the most universal problems, while the most refined method (experiment) has its own particular difficulties.

NATURALISTIC OBSERVATION: PROBLEMS OF OBSERVER INTERFERENCE AND IMPERFECT RECORDING

Observation is the most basic method of gathering scientific evidence, and it raises the most basic questions about validity—questions that arise in all methods of empirical research. We can see these questions by looking closely at the definition of naturalistic observation, which has two main requirements:

1. Complete and accurate recording of the relevant events
2. Minimal interference with the events

In practice, neither of these requirements can be completely met: researchers can only strive to approach them as ideals. The ways they fall short of the ideals open the possibility for alternative interpretations of observations. Let's look first at the ideal of minimal interference with events.

Problems Caused by the Presence of an Observer

Naturalistic observation strives to introduce "minimal interference with events," but there is no way to know for certain how much the observer's presence has changed things. To find out, we would have to observe the events both with and without the observer, and see how much difference there is. This is, of course, a logical impossibility and therefore we can never be sure how much the research process has changed the people and events being studied. The methods of naturalistic observation have raised this issue, yet it is obviously crucial to all methods of social science research.

The presence of an observer can affect observations in many ways. We divide them into two categories. When people behave differently because of a desire to create some kind of impression on the observer, we refer to these temporary changes as "on stage effects" (Agnew & Pyke, 1969) to suggest that people are acting for the benefit of an audience. Observers can also cause more persistent changes in the people and events they are observing that may continue even when the observer leaves the scene. We discuss these types of observer-produced effects separately.

"On Stage" Effects Experience has taught social scientists to identify situations in which the research process is most likely to interfere with events. One type of effect an observer can produce by merely being present has been called the *"on stage effect"* (Agnew & Pyke, 1969). This theatrical metaphor suggests that people may begin to "act" when they are aware there is an "audience." The problem of "putting on an act" can be expected to become more serious the more aware people are that there is an audience, the better they know what about them is being observed, and the more the subject of observation is personal or controversial. That is, the more difference it makes to people what impression they make, the more likely they are to act for the researcher. Below are some classic types of "on stage effects" and some methods used to control them.

Social desirability. People sometimes tell an observer what they think they "should" say. When people are asked about their values, many tend to report culturally acceptable values, even when they do not hold them. Such people's responses are influenced by their perceptions of social desirability. When people's adherence to a social norm is observed, it is reasonable to assume that the observer's presence may increase *apparent* conformity.

Evaluation apprehension. Sometimes people believe the observer to be somehow judging their personal adequacy or mental health. This belief, called evaluation apprehension (Rosenberg, 1965, 1969) obviously becomes stronger when the observer is labeled "psychologist." The effects of evaluation apprehension depend on the subject's perception of what mentally healthy people are supposed to do in the situation being studied.

Looking bad. Subjects of research occasionally try to make themselves look bad. This is perhaps due to a desire to sabotage the research or because the person feels something can be gained by looking bad. Some mental patients have been seen to do this when they fear being released from a comfortable hospital stay (Braginsky & Braginsky, 1967).

Demand characteristics. People sometimes try to please a researcher by doing what they think she/he wants them to do. Someone who means to please may become attuned to subtle cues in the interaction, called demand characteristics (Orne, 1962), that give a clue to what the researcher is looking for. Orne originally argued that subjects could be expected to accept these cues and would try to do whatever they thought the researcher wanted. However, subjects might also use these cues to "sabo-

tage" the study, or to outwit the researcher. There is evidence that this is a common attitude among people coerced into being research subjects, such as students who become subjects to fulfill a course requirement (Cox & Supprelle, 1971).

These examples suggest only a few of many types of on stage effects. People's behavior may alter in many different ways, depending on their beliefs about a researcher's identity or purposes. Ethnographers who study small communities may find that their hosts believe they are present to solve the community's problems, to critique the community, as government spies, as police informers, or in other roles that may have no relation to the researcher's actual purposes but that may lead the people being observed to behave in unusual ways because of the impression they want to create on the researcher. Martyn Hammersley and Paul Atkinson (1983) describe numerous examples from community studies, including this extreme case reported by Den Hollander (1967):

> In a town in southern Georgia [in 1932] it was rumoured after a few days that I was a scout for a rayon concern and might help to get a rayon industry established in the town. My denial reinforced the rumour, everyone tried to convince me of the excellent qualities of the town and its population—the observer had turned inito a fairy godmother and serious work was no longer possible. Departure was the only solution.

Such "on stage effects" are called *artifacts* of research, because they are created by the researcher and are not normally part of the phenomenon the researcher wants to study. Thus, to the extent that people are acting differently because they are "on stage," any observations of variables in their behavior are also measuring extraneous variables. These extraneous variables—the desire to look "healthy," to please or outwit the experimenter, to say the acceptable thing, and so on—provide possible alternative explanations for observed behavior whenever there is reason to suspect that people are "acting."

The on stage type of artifact is produced when people are aware that they are being observed, and when they desire to make some sort of impression on the observer. This is most likely to occur under these conditions:

> When there is little purpose for the researcher's presence other than to observe the subject—that is, when the observer is *obtrusive*. This is frequently the case in survey research, where subjects not only know they are being observed, but usually know what about them is being observed because the questions are straightforward.
>
> When the researcher holds higher status than the subject. If the researcher holds higher status this should increase the subject's desire to influence the impression he/she makes. The problem is most serious when the researcher can control important events in the subject's life, such as when a teacher or professor studies a student, or a psychiatrist or psychologist observes a mental patient, or a corrections staff member studies a prison inmate.

Observers are obtrusive whenever they are strikingly different from the people being observed and therefore particularly difficult for them to ignore. For example, in many anthropological and sociological studies the observer comes from a culture or subculture alien to those being observed, is markedly different in language, ethnicity, or social background, and does not know the norms of behavior in the group.

It is easy to imagine people in the presence of such a complete outsider trying to impress, or confuse, or play tricks on the observer. And it is easy to imagine the outsider not realizing what is happening, and failing to recognize that he or she is observing people on stage. So the possibility that people are "acting" provides a wealth of possible explanations of any behavior in the presence of an outside observer.

This possibility is a major threat to validity in research conducted by outside observers. Yet it is virtually impossible to conduct observations of some groups without bringing in complete outsiders. For example, it is almost always outsiders who want to conduct research on indigenous peoples of the Amazon, the workings of organized crime, and the play of small children. And with many other groups and activities, the researchers are frequently outsiders. So addressing on-stage effects is fundamental to the methods of ethnography and participant observation that are common in anthropology and sociology.

METHODS OF CONTROL

Here are some methods social researchers use to handle "on stage" effects.

Unobtrusive measures. Webb, Campbell, Schwartz, and Sechrest (1966) wrote a book on ways to measure subjects' behavior without their knowing it is being measured. These unobtrusive measures may or may not involve invasions of individual privacy. Consider these examples: to compare the popularity of various exhibits at a museum, the carpets in each gallery are examined for wear. To measure the effect of social status as an inhibitor of aggression, Doob and Gross (1968) had either a late-model Chrysler or an old, inexpensive car model stop at a light and stay stopped when the light turned green. The length of time it took the car behind to honk measured the inhibition of aggression. To measure racial prejudice, two people claimed to be identifiable by voice as black and white, dialed telephone numbers (ostensibly wrong numbers). The callers explained they were calling from a pay phone on the parkway, where their car had broken down. They were trying to reach a garage and had run out of coins. The people answering the phone were asked to please call the garage with the message. The number given was that of a researcher, who simply tabulated results (Gaertner & Bickman, 1972).

Deception. These last-mentioned unobtrusive measures also involve deception. On stage effects can be controlled by deceiving the subjects concerning the purpose, or even the presence, of the researcher. Thus, any attempt to respond to the researcher's purpose is nullified. Holdaway (1982) was a police officer who, after studying sociology, wanted to do observational research on the police. He decided to conduct the research covertly because he believed that if he had asked permission, the officers in charge would have denied permission or obstructed the research. Festinger, Riecken, and Schachter (1956) did a classic observational study of an apocalyptic religious group in which they wanted to test their hypotheses about how the group members would respond when the world did not end when they expected it to. They joined the group and did not reveal themselves as researchers, because to do so would have invalidated their observations or gotten them thrown out of the group. Researchers sometimes conceal part of their purpose or misinform their subjects deliberately in order to get more honest answers to questions.

Of course, there are some serious problems with deception as a strategy. For one thing, its ethics are questionable. There is a serious debate, especially among psychologists, about when deception is ever justified in research, and some guidelines have been developed (American Psychological Association, 1982). It is generally agreed, at the very least, that deception should be avoided whenever it is possible to get acceptable data by any other strategy. Many also feel that it is better to give up on some research questions rather than deceive participants in the research. A second major problem with deception is a practical one. Since it is known that social researchers, particularly psychologists, use deception, potential subjects are sometimes suspicious even of research that involves no deception. Thus, subjects' expectations to be deceived may influence their behavior.

Demand characteristics control group. One way to control any artifact in research is to manipulate it experimentally, using a comparison group design. One group gets whatever demand characteristics are in the experiment as planned, and another group gets a different demand, intentionally produced. For example, in a study on persuasion, it is desirable to be sure that any effects result from the persuasive communication used in the study, not the subject's desire to please the speaker, or some other extraneous variable. To control for this possibility, an investigator might run one group in which the persuader is introduced in the usual way, and another group in which subjects are also told that the experimenter disagrees with the point of view about to be presented. In this second group (control group), demand characteristics are *added,* to counter the persuasion attempt. Comparing this group with the experimental group will help determine whether demand characteristics influence persuasion in the experiment. An *evaluation apprehension control group* can be set up along similar lines to control for this extraneous variable.

Special controls for social desirability. In research that collects data by interview or questionnaire techniques, it is possible to control for the social desirability effect by the use of carefully worded questions. If, for example, people are asked to choose between alternatives that have been previously rated as equal in social desirability, their choice must be based on the content of the questions, rather than on the social desirability of the answers.

Inside observers. Observations by insiders—members of the group being observed—can be freer of on stage effects because the people being observed do not change their behavior for the benefit of the observer. This is one reason ethnographers often rely on informants—members of the group being studied who report to the researcher about what goes on in the group. The use of informants is not a form of naturalistic observation because it cannot hope to achieve complete and accurate recording of events as they occur. It is therefore vulnerable to other problems, such as biased observation or reporting by the informants and the possibility that the informant sees only a slice of life in his or her group and therefore gives the researcher a mistaken view. To get around these problems, social scientists sometimes train insiders in observational techniques. This procedure can help eliminate on stage effects without sacrificing completeness of observation. But it is not necessarily the case that insiders' observations are more valid than outsiders'. Insiders have an advantage in that they are more likely to know what is meaningful in a group, but they may also be so immersed in the group's culture as not to notice important aspects of life in the group that quickly strike an outsider.

Extended periods of observation. Anthropological and sociological field researchers, as well as ethologists who observe animal behavior in natural settings, typically spend long periods in observation before they produce a final report of their observations. They do this in part to control on stage effects by letting them dissipate over time: it is unlikely that the people or animals they are watching can maintain the same "act" for months or years.

Cross-checking observations. Another advantage of extended observation is that when the observer has seen a variety of behavior over time, it becomes possible to check the validity of the observations by comparing earlier observations with later ones, checking observations of some people against observations of others, and comparing different methods of observation (for example, comparing informants' reports with the researcher's own observations). These sorts of comparisons allow a researcher to distinguish behavior "on stage" from other behavior.

Many of the above methods of control assume that behavior "on stage" is somehow less valid than other behavior. But there is another way to think about social performances. Some social scientists believe with Erving Goffman (1959) and Shakespeare that all the world's a stage and that everyday life is like a play, in which people normally perform social roles for each other. From that point of view, it is possible to separate the roles people play for the researcher's benefit from the roles they play for the other people in their social settings. An observer might choose to record only the behavior he or she considers to represent the "natural" behavior of the observed, but it might be better to record both the behavior judged to be natural and the behavior judged to be "on stage," along with the reasons for making those judgments. This allows others to decide whether or not to accept the observer's judgment. And, if life is in fact best interpreted as a series of performances, then observations of people's performances in different situations is the best way to gain understanding, and performances for the researcher are as valid a slice of life as performances for others.

More Persistent Changes Caused by Research While on stage effects are serious problems for social research, the presence of a researcher can create more subtle and pervasive changes in the people being studied. This presence can, in some situations, cause people to change in ways that are more than just acting—that is, changes may occur that persist even when the subject is "off stage." Here are some classic examples.

Hawthorne effect. A famous set of experiments on worker productivity in an industrial plant in Hawthorne, Illinois, called attention to one possible on stage effect. The researchers (Roethlisberger & Dickson, 1939) reported that productivity increased every time the workers were shifted to new, experimental conditions, but soon leveled off, only to increase as soon as they were shifted again—even if they were shifted to conditions in which they had produced more slowly before. This behavioral pattern of improved performance because of the researcher's presence, which came to be known as the Hawthorne effect, has been attributed to the subjects' awareness that they were in an experiment, or that they were being given special treatment. Some have questioned the existence of the Hawthorne effect. Subsequent researchers have reexamined the data from the Hawthorne experiments and claimed that the Hawthorne effect never occurred there (Adair, 1984; Jones, 1992); questions

have also been raised on the basis of a number of studies about whether Hawthorne effects occur in educational settings (Adair, Sharpe, & Huynh, 1989). Nevertheless, it still seems plausible that Hawthorne-type effects may occur in some settings, such as when research subjects are suffering from boredom or lack social contacts (e.g., chronic mental patients, residents of nursing homes or schools for the retarded, and possibly even assembly-line workers like those in the original experiments).

Placebo effect. When a person expects a treatment or experience to change her/him, the person often changes, even when the "treatment" is known to be an inert or ineffective one. This effect is best known in research on drugs, in which the effect of the drug must be carefully separated from the effect of the fact that the patient is being given a prescription by a competent doctor. The "bedside manner" or the "power of suggestion" can heal too. This placebo effect has been offered as an explanation or partial explanation of voodoo death, religious healing, and psychotherapeutic cures.

Researcher expectancy effect. Robert Rosenthal (1966) had people look at photographs and judge how successful the people in the photos appeared to be. The experimenters in Rosenthal's studies were told either that the mean rating of success would be about +5 or about −5 on a scale of −10 to +10. The experimenters who were given the positive expectancy obtained more positive ratings from their subjects than the experimenters given the negative expectancy. Rosenthal suggested that the researcher's expectancy may somehow change her/his behavior toward subjects, and that subjects may respond to these subtle cues, creating a self-fulfilling prophecy: the researcher's actions cause subjects to behave as expected.

The most famous example of this effect comes from experiments in which grade-school teachers are told that certain of their pupils (randomly selected) have been tested and found to be "late bloomers" who can be expected to show great improvements in performance during the coming school year. At the end of the year, those students had in fact blossomed, as measured by such indicators as increased IQ scores, compared with pupils who had not been labeled late bloomers. Following the famous story of the street urchin who was taught to become a lady, this effect of a positive expectancy has come to be known as the *Pygmalion effect* (Rosenthal & Jacobson, 1968/1989).

There is some evidence that one way a researcher can communicate an expectancy is through the tone of voice in which instructions are given (Duncan & Rosenthal, 1968; Duncan, Rosenberg, & Finkelstein, 1969). In Rosenthal's study, experimenters with positive expectancies tended, for example, to emphasize "plus ten" more than "minus ten" when reading the instructions that described the rating scale.

Personal relationship effect. Sidney Jourard (1971) demonstrated that time spent in mutual self-disclosure of personal material by subject and experimenter could affect the rate of learning of meaningless material by subjects. It seems that subjects' performance may be affected by their emotional reaction to an experimenter as a person. Jourard suggests further that in typical laboratory experiments, in which the experimenter attempts to be impersonal (to "control" emotional reactions), subjects may act in an atypical manner. If so, people's behavior under these conditions may be, in part, a response to the extraneous variable of a "cold" researcher. This source of distortion can be called a personal relationship effect.

Personal relationship effects become increasingly difficult to avoid the longer a researcher spends making observations. They pose especially difficult problems in anthropological and sociological field research because over a long period of presence in a social group, many personal relationships are bound to develop and some of them are likely to change some of the individuals being observed, or possibly the entire group. In fact, the ability of an outsider to observe for any extended period depends on some sort of a personal relationship between the observer and at least one member of the group, who provides access to the group and is available to explain to other group members why the researcher is present. Decisions about which observations to record often depend on personal relationships, too. Imagine an anthropologist visiting a small, remote tribe of people previously unknown to Western societies. It may be impossible for such an observer even to tell which behavior is worth recording without learning the group's language and establishing personal relationships with several group members who can help explain the meanings of the group's activities. These "informants" are likely to want to learn about the observer, too—but that part of the personal relationship introduces group members to the mind of an outsider and inevitably introduces a new element into the group's life. The group may be changed permanently by its contact with the observer. Even when the observer and the observed are not so extremely different, these sorts of interactions still occur. It is impossible to remain in extended contact with other human beings without establishing relationships with them: the attempt to avoid relationships would be so bizarre as to be a relationship of its own kind, possibly one that is highly disruptive to what is being observed. For this reason, the possibility that the observer has, by the mere act of observation, changed what is being observed is an ever-present issue in observational research.

All these persistent changes that might be caused by the researcher's presence, like those that happen only when those being observed are "on stage," result from some extraneous variable unintentionally introduced into the research—the element of novelty, a person's expectation of change, the researcher's expectation about the behavior of the observed, or the quality of the relationship between the researcher and the people being observed. Each of these variables, whenever it may be operating, suggests an alternative explanation of behavior. It should be clear from the examples that these extraneous variables can affect any kind of observation—not just naturalistic.

METHODS OF CONTROL

Here are some methods used to achieve some control over these threats to validity caused by the fact of observation:

Blind measurement. To address the possibility that a researcher may affect the behavior of someone being observed by conveying subtle cues that communicate the researcher's expectancy, techniques of blind measurement can be used. There are two types of blind measurement. In one, the person being studied is kept "blind" to the researcher's presence—for example, by recording responses on paper or electronic media. In the other, the researcher is kept "blind" by having some other member of the research team, who is not told what to expect, interact with the person being observed and make the observations.

Double-blind technique. This is an extension of the blind measurement technique.

In a double-blind experiment, both the researcher *and* the subject are blind to the treatment (i.e., they do not know what treatment the subject is getting). This method was developed for drug research, but it also has applications in social science. In a drug study, one experimenter assigns subjects to treatment conditions, and prepares the medication for all subjects. All preparations look, smell, and taste alike, although the contents are different. A second experimenter then administers the drugs to all subjects, without knowing who is getting what. This procedure controls for self-fulfilling prophecies by giving the researcher in contact with the subject the same expectancy for all subjects. It controls for the placebo effect by giving all subjects the same expectations of help. The effect of personal relationship is probably about the same for both treatment groups.

Placebo control group. This methodology was also developed for drug research, and it too has social science applications. A placebo control group is treated exactly as an experimental group, except that instead of the experimental drug, a substitute is used that is physiologically inert (has no physical effects) but is indistinguishable from the drug by sight, smell, or taste. The purpose of the placebo is to separate the effect of the drug from the effect of expecting to be cured, talking to the doctor, and other aspects of the treatment situation that might help the patient, but do not depend on the specific medication given. Thus, any improvement in the drug treatment group above and beyond what is observed in the placebo group can be attributed to the drug. This procedure controls for the Hawthorne effect and the self-fulfilling prophecy, which both depend on the situation surrounding administration of treatment, rather than the specific treatment itself.

The principle of the placebo control procedure can be used in a variety of settings. In research on psychotherapeutic techniques, teaching methods, and treatment programs for juvenile delinquents, drug addicts, and so on, there is probably no completely inert treatment. In such research, various comparison groups have been used—people who want psychotherapy but are on the waiting list at the clinic, people getting an established, nonexperimental form of treatment, people meeting in a discussion group not designed as treatment, and so forth. While people in all these comparison groups might undergo change as a result of their treatment, no group is, strictly speaking, a placebo group. However, these comparison groups have the same function as a placebo group because they represent treatments that either are presumed to be relatively inert, or at least do not have the specific effects expected from the experimental treatment.

Warm-up period. One way to minimize the Hawthorne effect is for the researcher to spend some time with the person or group being observed to diminish the novelty of the interaction, which is believed to be a cause of the effect. A warm-up period might also control for personal relationship effects caused by researchers trying to be impersonal. However, it might be that the researcher forms strong relationships with some of the people being observed and not others, introducing an extraneous variable. This possibility suggests another control tactic.

The "canned" researcher. This is an invented name for a commonly used method of controlling demand characteristics and personal relationship effects. Unlike "blind" observers, who interact with the people being observed in ignorance of the researcher's expectation, "canned" observers are automated. This control technique deals with difficulties in the researcher-subject relationship by eliminating it. Any

instructions to be given to the research subject might be written out or presented in prerecorded form. (Note that while this procedure holds the researcher-subject relationship constant, it does not meet Jourard's (1971) criticism that "cold" experiments bring out atypical behavior in subjects.) If the subject never meets the experimenter personally, it is very difficult for the researcher to communicate expectancies by means of subtle nonverbal cues. Thus, the effects of self-fulfilling prophecy and demand characteristics are lessened. Also, it is certain that all subjects have received the same instructions, even down to tone of voice.

MAKING VALID OBSERVATIONS WHEN CONTROL TECHNIQUES CANNOT BE USED

The above techniques for controlling the research situation are intended to prevent the research act from changing what is being studied. But using such methods of control is not always possible in social science research. The difficulties are most obvious for field researchers who spend long periods in foreign cultures or other unfamiliar social settings. All the above techniques with the exception of the warm-up period are impossible to implement in such settings. Good field researchers are sensitive to the fact that whatever they do has the potential to permanently alter the setting they are studying. Even laboratory researchers must deal with this fact: if subjects respond differently to a canned researcher and a real researcher, how is one to know for certain which response (if either) corresponds with what they would have done in the researcher's absence?

Ethnographers have given considerable thought to the problem they call *reflexivity:* that social researchers are part of the world they study so that to some degree their observations are always observations of themselves and their effects on their surroundings. A good detailed account of the implications of reflexivity for the practice of field research is given in Hammersley and Atkinson's (1983) book, *Ethnography: Principles in Practice.* Reflexivity implies that there is no way "in principle . . . to isolate a body of data uncontaminated by the researcher" (Hammersley & Atkinson, 1983, p. 14). Reflexivity further implies that no observation is free of threats to its validity and that the only way to reach valid conclusions from observation is ultimately to study (rather than try to eliminate) the researcher's effects on what is being observed. In this view, observer effects, rather than being "artifacts" to be eliminated by control, are data to analyze. The strategy of *cross-checking observations,* already mentioned as a way to address on stage effects, is the best way to understand permanent changes caused by the researcher's actions. Even if no observation is fully valid, useful knowledge can come from examining how different kinds of observations distort reality in different ways. By comparing different kinds of observations, it is possible to learn what is invariant in the face of different kinds of interventions from outside (the researcher's activities) and also how the people being studied respond differently to different kinds of interventions. Both of these are important information on whatever is being studied.

This is not to say that whatever a researcher does in the field provides equally valuable information. Ethnographers carefully consider what impression they want to give to the people they are observing to increase their chances of getting useful information. We have already mentioned situations in which researchers have had to

conceal their identities as social scientists because being open would have made it impossible to make valid observations. Field researchers may carefully choose their clothing so as not to stand out too much from the people they are studying. Sometimes they are very careful how they present their own beliefs and attitudes: sociologists who study deviant groups find that group members often want reassurance that the researcher does not disapprove of them. Such choices about self-presentation may well affect the quality of a researcher's observations; certainly, experienced ethnographers claim that they do.

Keeping distant or getting close? Some researchers prefer to present themselves very much as outsiders, maintaining a strong degree of social detachment and reserve from the people they are observing and justifying this stance on the ground that social distance is needed to maintain objectivity. Others prefer to establish close personal relationships with members of the group being observed and justify this stance on the ground that deeper understanding is possible when emotions are allowed to enter the social interaction. A good case can be made for either approach, but a researcher must choose some position on the continuum from pure observer to pure participant, understanding that each position affects the observations. A common piece of advice is for the field researcher to cultivate the role of an "acceptable incompetent" (Lofland and Lofland, 1984), someone who will not be rejected by the group, and who may also benefit from instruction by group members in the way the group works. But certainly, people behave differently around incompetents in their culture than around sophisticates, so even the acceptable incompetent's position affects what is observed.

The question of what stance to take in making observations highlights a dilemma that is very stark in long-term field research but that exists to some degree with observations generally. Each stance holds the researcher's attitude constant, so it is a kind of control, but each one also affects what is observed, so it is also a kind of extraneous variable. To make progress despite this dilemma, in which every control is a source of bias in the observations, good field researchers often try different ways of interacting in the field setting as a way to cross-check their findings. They also are explicit about how they presented themselves to the people they studied, so that others can make their own judgments about whether the researcher's behavior gives reasons to question the researcher's findings.

Comparing observers. One problem can never be solved by a researcher observing for a long time or cross-checking his or her own observations. Every researcher is a particular individual with a particular social background, style of interaction, set of preconceptions, and so forth, so all the interactions of that individual with the people being studied may be colored by the characteristics of that individual. Gender differences provide a good example. Hammersley and Atkinson (1983) point out that "in male-dominated settings, for instance, women may come up against the male 'fraternity,' from which they are excluded; women may also find themselves the object of 'hustling' from male hosts . . . [but] female researchers may find advantageous trade-offs. The 'hustling' informant who is trying to impress the researcher may prove particularly forthcoming to her . . ." (p. 85). We are not advocating the use of sexuality as a research tool, only warning that sex and gender may affect research observations in various ways. Male researchers, of course, may also affect the behavior of the people they study merely because of their gender, and equally significant effects can result from the researcher characteristics other than gender.

Ultimately, the best way to find out whether the researcher's individuality made

a difference is to compare the reports of different observers. It may be especially useful, depending on the situation, to compare the observations of researchers who differ in cultural background, gender, race, religion, social status, and in their strategy of distancing themselves from, or interacting closely with, the people they study.

To summarize a long discussion: Researchers can produce two types of unwanted effects on what they are observing by their mere presence: on stage effects and the more persistent "real" changes in people that can result from the research process. These unwanted effects exist because researchers unintentionally introduce extraneous variables when they observe events. Table 3.1 presents the material in this section in condensed form.

This discussion and the tables in this chapter are intended to help you to raise questions when you read scientific literature. If you have a good sense of how the research process can change people and events, you will be in a position to offer plausible alternative explanations for the findings of some of the research reports you read. Only when all reasonable explanations are collected can you make an educated judgment about how strongly a set of research results justifies an author's conclusions.

Table 3.1. Extraneous Variables Due to the Presence of an Observer

Extraneous Variables	Alternative Explanations	When a Problem	Methods of Control
	On-Stage Effects		
Social desirability	Subject may be saying what he/she "should" believe	Survey research, controversial topics	Careful construction of questions; unobtrusive measurement; extended observation
Evaluation apprehension	Subject may be trying to impress someone juding "mental health," IQ, etc.	Survey research; when researcher has high status	Deception; unobtrusive measurement; extended observation; insider observers; comparing observers
Faking bad	Subject may be trying to sabotage research	High status researcher; coerced subjects	Deception; unobtrusive measurement; extended observation; insider observers; comparing observers
Demand characteristics	Subject may be doing what he/she thinks researcher wants	High status researcher; volunteer subjects	Decpetion; unobtrusive measurement; special control group; extended observation; "canned" researcher; insider observers; comparing observers

(*continued*)

Table 3.1. *(continued)*

Extraneous Variables	Alternative Explanations	When a Problem	Methods of Control
	More Persistent Changes Caused by Research		
Hawthorne effect	Performance improves merely because of change in routine	Subjects lack social contacts	Comparison group with different treatment; warm-up period
Placebo effect	Subject may be changing because he/she expected to	"Therapy" settings where people expect to change	Warm-up periods; placebo control group; double-blind technique; "canned" researcher
Researcher expectancy (self-fulfilling prophecy)	Researcher may subtly communicate an expectancy that subject acts to fulfill	Researcher and subject in close contact	Bllind measurement; double-blind technique; placebo control group; "canned" researcher; deception about expectancy
Personal relationship effect	Subjects may perform differently because of nature of relationship with researcher	A general problem	Warm-up period; "canned" researcher; comparison group with different relationship; comparing observers
Reflexivity problems	Responses may be due to researcher's personal characteristics or behavior with subjects	General; becomes more serious with more researcher presence	Comparing observers; cross-checking results

Problems of Incomplete or Inaccurate Recording

By our definition, naturalistic observation requires "complete and accurate recording of the relevant events." Like noninterference with events, this is an ideal rather than a realistic possibility in social research. One problem is that observers may be excluded from observing certain events that are essential for understanding the people being observed. Even when access is unlimited, a researcher cannot be sure about the selection of relevant events and the rejection of irrelevant ones. Complete recording is typically a practical impossibility because too much may be going on, even in a simple social situation, and some further selection may have to be made. And it is always possible for a researcher to record inaccurately without realizing it.

Selection and potential inaccuracy are always a part of naturalistic observation, as with all scientific methods. They pose a difficult question for an observer: How can I know that nothing important was left out or distorted in my record of observations? And they pose a doubly difficult question for readers of research: If the observer omitted important events or distorted the observations—and may not even realize it—how can I tell?

Incomplete Access Observers often have difficulty gaining access to the phenomena they want to observe. Hammersley and Atkinson (1983) discuss two illustrative examples. Chambliss (1975) reports his difficulties trying to study organized crime in Seattle. First, dressed like a truck driver, he visited a skid-row cafe and learned about an illegal poker game in the back room. Over some months, he played poker, visited pornography shops, and engaged in conversations with gamblers, bartenders, and all sorts of low-status participants in organized crime, but never gained any understanding of how the criminal enterprises were organized. Finally, after revealing himself to the manager of the cardroom as a sociology professor with a "purely scientific" interest in the operation, he began to receive calls from others at higher levels in organized crime who were willing to talk with him. He became able to see, if only at second hand, things he could never have seen without the assistance of others.

Hansen (1977), in studying rural village life in Catalonia, politely asked villagers for interviews, and learned very little. By chance he interviewed one of the few noblemen in the area, who told him that as a person whose looks and education marked him as superior to most villagers, he should command people to give interviews. The count then accompanied Hansen to visit landholders and ordered them to give Hansen all the information he wanted. After that, it became fashionable to be interviewed, and Hansen reported a flood of volunteers.

In both these examples, researchers got assistance from helpful insiders who gave them access to observations and other information they could not otherwise have obtained. Their reports would have been much different, and probably much less insightful, without this enhanced access. It is of course possible that the researchers were still misled somehow by their informants, but because they had greater access, they were put in a position that allowed them to compare what they learned from different informants and thus offer a more accurate picture of what they were observing than they could hope to write with more limited access.

It is always wise to ask of observational studies whether the researcher had access to all the important aspects of the phenomenon being observed. Some of the same researcher characteristics that lead to on stage effects may also lead people to conceal important information from researchers. Most social groups have secrets, and even conceal some social phenomena from some individuals within the group (children, for example). They often conceal the same information from researchers, especially when their scientific role is obvious or when they are strikingly different from the people being studied in gender, race, social status, or other important characteristics. An observer with incomplete access cannot record all the relevant events, and the selection of events created by incomplete access will usually tend to mislead the researcher in some way and to bias the researcher's interpretation of what was observed. This sort of bias comes from the effort of those being studied to conceal or deceive.

Researcher Selectivity Researchers can also introduce biases of their own into their observations. This can happen when a researcher begins with a theory (often implicit) that directs the questions that will be asked and the phenomena that will be observed. It can also happen when factors in the researcher's background or some aspect of the way the observer enters the situation leads to increased attention to parts of a phenomenon, which are selected as relevant, and a failure even to notice other parts. Consider the following three examples:

Suppose an education professor studies parent participation in education by observing an elementary school in a middle-class African-American suburb where most of the students are performing well below grade level. The professor notices that parent visits to teachers and the principal are very infrequent, and interprets this observation as evidence of parent disinterest in the children's education. Such a conclusion would fit well with "cultural deficit" theories about the causes of poor school performance among African-Americans and with concepts and theories that emphasize parent interest and involvement as an important factor in pupils' performance. But a study emanating from the experiences and perspective of African-American parents in one such school told a different story. African-American parents' attempts to participate had been systematically discouraged by school officials, who refused to make appointments, broke appointments without reason, and treated interested parents dismissively and with disrespect (S. Stern, 1994). What looks like parent disinterest from a dominant perspective, such as that of most white, middle-class researchers or of school officials, looks like a quite different phenomenon—parent push-out—when observed from the parents' position in the system. As the parents see it, they stay away not out of disinterest, but because they are unwelcome and because their visits accomplish nothing positive for their children. What is significant here is not just that an observer can reach the wrong conclusion about what is going on, but that the observation itself may be biased: the observer fails entirely to observe an important part of the process (the part the parents see), and does not realize it.

Several best-selling guidebooks purport to offer applicants to American colleges all the important information they need to choose a college. Their authors collect data on the colleges, and some even send observers to describe the quality of life on campus. The guides report all sorts of information, from the number of books in the library to the availability of vegetarian dining to the quality of the college newspaper, but they do not report on many things of interest to African-American college applicants, such as the graduation rate for African-Americans, the number of black professors, the availability of programs for students who need remedial work, and the climate of race relations on campus. To fill the gap, *The Black Students' Guide to Colleges* (Beckham, 1984) has been published and updated. The success of this volume suggests that the best-sellers are really *white* students' guides, without saying so. They report a great variety of information, but their observations of the colleges are incomplete because they omit information that is essential to many potential students. Some of what they omit is important for *all* races of students, such as the information about race relations and remedial programs. Although the best-sellers claim to be based on observations of everything important about a college, they are not, and again, the authors do not realize it.

Some feminist authors (Dalmiya & Alcoff, 1993; Ehrenreich & English, 1973) claim that when physicians displaced midwives as the main practitioners of obstetrics, much knowledge was lost because the physicians (who were almost exclusively male) did not fully observe, and therefore did not fully understand, childbirth. Physicians usually attended women only at delivery, whereas midwives attended them throughout labor, and physicians considered only physical matters of childbirth (and then, only part of the body), whereas midwives also concerned themselves with psychological aspects. And because physicians were almost exclusively male and midwives female, there were stark differences in their abilities to understand childbirth empathically and to be sensitive to subtle changes in the condition of a woman in labor. The feminists claim that physicians, because they lacked the perspective and knowledge that midwives had, were inferior observers of their patients and inferior practitioners. One example given in this literature is that physicians innovated the practice of giving birth from a supine position, which allowed the obstetrician greater control but had no advantage for the mothers' or babies' well-being and, as midwives knew, made delivery more difficult for the mothers than it was from a sitting position.

In each of these examples, well-meaning observers who want their observations to be complete and accurate fail to see all of what they are trying to observe and, as a result, the knowledge they develop is faulty. Selective observation may have many causes. It may come from theory or other implicit presuppositions, such as that schools are equally open to parental involvement of African-American and white parents, or that African-American parents are culturally uninterested in their children's education. It may come from failure to recognize that not everyone shares one's social position and attendant concerns—the failure to discuss the climate of race relations in college guides may be an example. It may come from the professionalization of the observer—for example, the physicians whose training and experience led them to focus only on physical aspects of birthing and more on delivery than on labor. It may come from lack of empathy rooted in differences between the observer and the observed: white observers of black parents and male observers of the birth process may fail to observe what someone with greater empathy would quickly see.

Whatever the cause, the effect of selective observation is not only that things are left out, but that they are left out systematically. As a result, the entire observation is biased or distorted, as well as incomplete.

Systematic neglect of certain information is almost a universal problem in social research, because all methods of observation involve choices about which data to observe and which to ignore. The difficulty with making selections is that we do not know whether every possible fact has an equal chance of being observed. With human observers, it is safe to assume that the facts do *not* have an equal chance of being observed, because people have theories, or at least mental sets to look for certain kinds of facts, and because observation is affected by the observer's position in relation to what is being observed. Often we, as observers, are not aware of the classes of information we are ignoring, and this is as true of scientists as it is of everyone else. When a researcher selects information to look at, the reader generally gets a *biased sample of information.*

METHODS OF CONTROL

One way to control this is not to try to eliminate the bias (how can one know when it is gone?) but to be explicit about it. Researchers do this when they identify variables and define them operationally. The reader then knows exactly which informa-

tion was observed (information about the variables mentioned) and which was left out (everything else). The selectivity is still there, but the bases for the selection are known to all. In short, the way to handle the inevitable bias resulting from a researcher's selection of events is to use a research method that selects variables and defines them operationally. For this reason, naturalistic observations and retrospective case studies tend to be done early in the research history of a subject, before enough is known to decide which variables to study in depth. Later in the research history, when *theories* are developed, it is easier to make the researcher's bias explicit. Theory dictates which events should be studied and which neglected; not all events are equally likely to be studied. Thus, a theory simplifies a researcher's job by defining some facts as irrelevant. A researcher with a theory need not observe everything, and can therefore be more careful about measuring what is considered most important. Theory is also valuable in that it makes explicit the bias that is inevitable whenever an observer chooses not to record everything. All this is in addition to the major values of theory: to advance understanding and give direction to research.

Another way to address the problem of researcher selectivity is to use one observer's biases to reveal another's. The African-American parents could see things the observer of the school could not; black observers of colleges saw things white observers overlooked; and midwives could see important aspects of childbirth that male physicians overlooked. This does not necessarily mean that African-Americans and midwives were the best observers—they may have overlooked important things, too. But comparing observers, and particularly comparing observers who can be expected to have different points of view on a phenomenon, can reveal each researcher's selectivities and biases, and thus lead to a more complete and accurate picture than any one observer is likely to produce, no matter how carefully that observer records events.

Although the above examples focus on observational methods, the same problems arise with other research methods as well. One reason is that all research methods involve observation, and whenever there is room for judgment in making or recording observations, biases can enter. Moreover, faulty observations lead to faulty hypotheses for other research methods. Consequently, researchers using other methods may head off in the wrong direction by paying attention to irrelevant variables or failing to study important ones.

Researcher Distortion The same factors that can affect what an observer notices can influence the researcher's interpretation of events. Researcher distortion is a serious problem, especially where strongly held values are at stake and where a researcher has a stake in a particular hypothesis or theory. The evidence of research on attitudes suggests that anyone who spends many years of effort working on a theory is likely to come to believe in it, and this may affect what he or she sees. Consider an example: A psychologist who does group psychotherapy professionally wants to assess the effectiveness of her therapy. She believes that a diversity of personalities among the therapy group is counterproductive. She evaluates the progress of a diverse group of patients seen together and another diverse group of patients she is seeing in individual psychotherapy (control group). At the end of therapy, she reviews her notes, and rates patients "much improved," "somewhat improved," "no

change," "somewhat worse," or "much worse," compared with when therapy began. Since she knows who was seen individually and who was seen in group, and she has a stake in the outcome, we might not want to trust her ratings as a measure of patient improvement. Suppose her observations were that the group patients did not improve (just as she expected). The psychologist might conclude that diversity in therapy groups is counterproductive, but we could offer an alternative explanation: Because of her bias, the psychologist did not see evidence of improvement in the group-treated patients and exaggerated the improvement of those in individual therapy. This distortion is most serious when the reliability of observations is questionable. However, even when variables are carefully operationalized, distortion is possible.

An observer's social position can also cause observations to become distorted. The example of parent participation in schools suggests how this can happen. A social scientist who comes from a privileged majority-group background may have a basic belief, rooted in personal experience, that social institutions are generally responsive to the needs of individuals. Such a researcher would probably conclude that when parents are failing to intervene in a school where their children are performing poorly, the cause must lie in the parents rather than the school. As a result, the researcher might do two things: fail to look closely at the parents' behavior and experience (selective observation), and misinterpret the lack of parental contact with the school as parental unconcern (distortion). A social scientist whose personal experience had included disdainful treatment by official representatives of social institutions might be less likely to engage in the same distortion.

METHODS OF CONTROL

If the psychologist conducting the study of group therapy is a conscientious scientist, she does not trust her own judgment, but brings in someone else to evaluate the patients. She would control the possible effects of her distortion by using blind measurement. With the judgment of a competent colleague who does not know her hypothesis, she can obtain more accurate information about each patient's progress. The judge would review a transcript of the therapist's notes, edited to remove information on whether the patient is being seen individually or in a group, and would make the same ratings the therapist might make. Whatever bias the colleague may have would not influence the results because this judge doesn't know which patients were seen individually and which in group, or that a difference between individual and group therapy is expected.

Control could go a step further. The researcher could *misinform* the judge about the hypothesis, or about her bias, or about the patients' progress (e.g., she could tell the judge that none of the patients seemed to be responding to treatment), and let the judge evaluate each patient. This procedure might be an improvement because it would counter any subtle communication of the researcher's bias that might prejudice the judge.

Distortions caused by social position can most easily be discovered and corrected by observers with a different social position. In research on racism, the observations of researchers of different racial backgrounds might complement each other. Differences between the observations could then be interpreted as possible omissions or distortions on the part of either observer, or both. Similarly, where gender is an issue

in what is being observed, having male and female observers can help correct distortions. When power is an issue, it can help to have observers who can take the point of view of those with power and of those without it.

To summarize this discussion of incomplete or inaccurate recording, observational data are often open to alternative explanations because of incomplete access, researcher selectivity, and distorted observation. Table 3.2 briefly summarizes the ways this can happen and some methods for controlling, or at least revealing and understanding, the limitations of observational research.

RETROSPECTIVE CASE STUDY: PROBLEMS OF MEMORY

Retrospective case studies share many of the limitations of observational research, particularly problems of access, selectivity, and distortion. In addition, they have one essential characteristic that allows for alternative explanations for their findings. Because they collect data from the past, retrospective case studies often rely on people's (faulty) memories. Reliance on memory is not unique to retrospective case studies, but the problems associated with memory appear most clearly in this research method.

In retrospective research, there is selectivity and distortion not only on the part of the researcher, but on the subject's part as well. What someone remembers is not only incomplete, it is systematically incomplete. Ordinary people, like scientists, have theories about the relationships between events, and what they judge as unimportant tends to be forgotten. These biases are usually unexpressed. Furthermore, memories can be distorted to fit the view that makes a person most comfortable at present. Suppose, for example, a researcher is interested in the predisposing factors in juvenile delinquency. A sample of delinquent boys is selected, and each is asked questions about his relationships with his parents. Most of the boys report that their fathers were frequently absent from the home and spent little time playing with them. Can their memories be trusted? It could well be that these delinquent boys are rebelling against their fathers' authority and are justifying their rebellion by remembering the times father was away and saying that father didn't care. It is hard to know what produced the boys' reports if their memories are the only evidence available. (This problem is not restricted to retrospective case studies—selective memory is a very serious problem in correlational research and even in experimental research when variables are measured by people's accounts of the past.) Memories are most likely to be distorted when distortion can be used to justify one's actions and maintain or enhance one's self-esteem.

Table 3.2. Extraneous Variables Due to Incomplete Access, Selection, and Distortion

Extraneous Variables	Alternative Explanations	When a Problem	Methods of Control
Incomplete access	Unseen events may explain what is observed	Researcher is an outsider; those observed want to conceal information	Involve insiders in the research
Researcher selectivity	Events are due to causes researcher's theory considers unimportant; to causes someone in researcher's social position can't see	Researcher and subjects are from very different social groups; research lacks operational definitions	Specify selectivity by operationalizing variables; compare observations by people from different social positions or using different theories
Researcher distortion or bias	Researcher's evaluation of data may be colored by preconceptions	Researcher knows hypothesis; has stake in results; comes from different social group from those observed	Blind judges; misled judges; compare different observers

METHODS OF CONTROL

The only ways to control the effects of selective and distorted memory involve using other sources of information. In a study of the causes of delinquency, for example, it is possible to ask the delinquent boy and his parents the same questions about the period of his childhood. This way the amount of distortion between different memories will be known, even if it is impossible to know whose memory was most accurate.

In some retrospective research, it is possible to rely on *archival records* that do not depend on memory. The Kerner Commission study of urban riots (Chapter 2) used records of incomes and unemployment rates to determine the economic conditions of cities where riots occurred.

Another approach is to use a research method that does not rely on memory. In the study of delinquency, for example, it is possible to do a *prospective study,* in which a large number of children is *directly observed* before some of them become delinquents, to identify the differences between the children who do and don't turn out delinquent. Another alternative is experimental research. In the study of delinquency, a group of young children can be given whatever delinquents are presumed to lack, to see if their rate of delinquency turns out to be lower than that of a control group.

Table 3.3 summarizes how selective or distorted memory can allow for alternative explanations of the results of studies that rely on memory.

SAMPLE STUDY: PROBLEMS OF OPERATIONAL DEFINITIONS AND GENERALIZING ABOUT POPULATIONS

Sample studies have three important characteristics not generally present in naturalistic observations and retrospective case studies.

1. Variables in sample studies are operationally defined.
2. Sample studies generalize about populations from information about samples.

Table 3.3. Extraneous Variables Due to Memory

Extraneous Variables	Alternative Explanations	When a Problem	Methods of Control
Selective or distorted memory (in subject)	Subject's memory may be distorted to fit his or her "theory" or current opinion	Retrospective research relying on memory; subject's self-esteem at stake	Compare two memories; use archival records; prospective research

3. Sample studies involve collecting the same information about a number of different people or events.

Since sample studies share these characteristics with correlational and experimental research, the validity problems that exist in sample studies are also present in the other forms of quantitative research.

Invalid Operational Definitions

Whenever operational definitions are used, the possibility exists that they are invalid. Unfortunately, there is no research design or procedure that will protect research from *invalid operational definitions*. In evaluating research, it is your job to think about the operational definitions used. Ask yourself if the definition of a variable being studied might be measuring something else. Here are some examples. Most intelligence tests require knowledge (often reading knowledge) of the language in which they are given—this means they are also measuring acquired language skills. Juvenile delinquency can be defined in terms of convictions in court, but convictions are more frequent when defendants don't have private counsel—this means that the definition of delinquency is also measuring economic status. If a researcher using this definition of delinquency discovers that delinquents are educationally deprived, the findings may only mean that poor people get poor educations. In general, when an operational definition measures more than one thing at a time, whatever is said of one of the variables could, with justification, be said of the other(s) as well. Operational definitions that measure more than one thing at a time are said to be *confounded*, and the variables measured together are said to be confounded with each other. When an operational definition is confounded, any conclusion drawn about the variable it is supposed to measure may just as well be drawn about the extraneous variable confounded with it. In evaluating research, keep constantly aware that operational definitions are not necessarily the same as the variables they are supposed to measure. If you should even suspect confounding, offer an alternative explanation based on the extraneous variable buried in the operational definition.

Biased Samples

There are some difficulties associated with drawing inferences about populations based on information about samples, and they can best be illustrated by an example. Suppose a researcher wants to determine the birth control practices of married couples in Vermont. There are, let us say, 120,000 such couples and, to make things easier, let us also assume a list of their names and addresses is available. Still, 120,000 is too many couples to survey, and so a sample of 500 couples is taken. Skipping the important details for now, let's say that of the couples surveyed, 27% use the pill, 19% use sterilization (either partner), 17% use intrauterine devices, 17% use diaphragms, condom, and/or foam, 5% use rhythm, and the remaining 15% use no birth control. Is it safe to say, for example, that the pill is the most commonly used method of birth control among Vermont couples? Not necessarily. Here are some problems.

Samples rarely contain exactly the same proportions of anything as the population from which they are drawn. (If you doubt this, toss a coin 100 times, and see if you get 50 tails. You will probably get *about* 50 tails, but you will probably not get exactly 50.) If the sample is a good one, it is fair to conclude that *about* 27% of the couples in the Vermont population use the pill, and that *about* 19% use sterilization. To decide whether the pill is more commonly used, you would need to test the hypothesis that "about 27%" in this sample is greater than "about 19%." To do this, you would have to determine the likelihood that, given a population in which the pill and sterilization are used with equal frequency, a sample of 500 couples would include 27% using the pill and 19% using sterilization. If this is highly *unlikely*, you can be reasonably sure that the pill and sterilization are *not* used with equal frequency in the whole population. There are statistical procedures to determine this likelihood. In general, the larger the sample, the more certain you can be that an observed difference corresponds to a difference in the whole population. (The more often you toss a loaded coin, the more certain you can be that it's loaded.) When samples of 500 are taken out of a population of 120,000, they are likely to be *somewhat* different from each other and from the population, and the generalizations you can make from a sample can be expected to be slightly inaccurate. The larger the sample, the less inaccurate it will be. This error of inference is called *sampling error*.

Sampling error is an unavoidable problem when a scientist attempts to make inferences about a population from less than complete data. However, it need not lead us to question the internal validity of research for two reasons. First, if the sample is chosen carefully, so as to be representative of the population, the amount of sampling error can be placed within known limits. You may see such statements as "We can say with 95% confidence that Dewey will win within 3% either way of 53% of the popular vote." This means that 5% of the time the error will be greater and that 2½% of the time Dewey will not win a majority. Although such an error may embarrass the prognosticator, it is predictable. We know how much error to expect how often.

Second, and more important, when a sample is representative of a population, the error is equally likely to go in either direction. Data from representative samples vary *randomly* around the data that would be collected from the population; the findings are not systematically distorted by the measurement of an extraneous variable. Thus, while sampling error does limit the certainty of any inference about a population, it is a weak argument for questioning the validity of a researcher's conclusions.

The above comments were all predicated on the assumption that the sample of 500 Vermont couples was *representative* of all couples in the state. The must serious problem with sampling involves being sure the sample is representative—that it is

not systematically distorted by some extraneous variable. Suppose, for example, that the Vermont sample included only half the proportion of Catholics that exist in the state population. Such a sample is not representative; it is *biased.* A *biased sample* is one that contains a *systematic error:* it is consistently different from the population in a particular direction. In the present example, a sample with few Catholics probably underestimates the proportion of couples using rhythm and no birth control. A biased sample is one that consistently misrepresents the population from which it was drawn; data from such a sample differ from population data in a particular direction because of the presence of an extraneous variable. Name the extraneous variable, and an alternative explanation follows: "Rhythm was found to be the least popular birth control method in Vermont because the sample *underrepresented Catholics,* not because it is least popular."

METHODS OF CONTROL

The only way to be certain that a sample is representative is to use a truly *random sample.* This presumes a complete list of the population (which was available), and a systematic procedure that allows everyone in the population an equal chance of being chosen for the sample. This might be done by putting all the couples' names and addresses in a computer, assigning each couple a number, and using a program that generates random numbers. The first 500 numbers that correspond to numbers that had been assigned to couples would determine the people sampled. While this procedure is possible, it might exhaust the researcher's budget to travel to the remote locations where all these people may live. Therefore, random samples, while they are theoretically ideal, are rarely used in large-scale sample studies.

A usual procedure is to choose a sample on some convenient basis, assuring that the sample is equivalent to the population with respect to several variables considered important to the research. For the Vermont birth control study, we might agree that it is important to make sure the sample and the population are similar in age distribution, religion, rural or urban residence, and number of children already born, since these factors probably influence birth control methods. If the researcher took a sample of people from rural areas, towns, and cities (in the same proportion as the state population), chose individuals in these locations on a random basis, and showed us that the sample was very close to the overall population of couples in terms of the other variables mentioned, we might be willing to accept the sample as representative. Other than true random sampling, there is no absolute rule for drawing a representative sample. All we can ask of a researcher is that the sample is representative of the population in those things that are probably relevant to the research question. If we are assured of that, we can proceed on the assumption that the error in this sample is no different from the error in a true random sample.

When evaluating a piece of research that uses sampling techniques, consider the population being sampled, and then think about the method used to draw the sample. If you can think of a way in which the sample may be systematically different from the population, ask yourself whether this bias could have influenced the results. If it could have, an alternative explanation is possible.

The above discussion has concerned the difficulties in concluding that data from a sample accurately represented the population *from which the sample was drawn.*

To ask whether the findings are true of other populations is to raise another question, which is discussed in Chapter 5. In the Vermont example, we have discussed whether one can be justified in drawing conclusions about adult married Vermonters. If one can, these conclusions still may not apply to the birth control practices of unmarried Vermonters or of people living elsewhere. We emphasize this point because researchers often take samples from much more restricted and less interesting populations than the adult married couples of Vermont. Most educational and psychological research uses conveniently available populations, such as "third grade pupils in the Horseheads Central School District" or "introductory psychology students at Moreland University, spring term, 1993." The question of internal validity often becomes the question of whether results hold true even for such restricted populations as these.

Sampling bias is possible even with such restricted populations. One common source of sampling bias is the use of volunteer subjects for experiments. Unlike the average person, the person who volunteers for psychological research is likely to want to please (demand characteristics), and may also be unusually well motivated to perform. Consequently, what is true of volunteer subjects may not be true of the population from which they are drawn. Thus, when volunteers are used, their desire to please (extraneous variable) may provide an alternative explanation of their behavior.

Uncontrolled Variation in Information

It may seem easy to collect the same information from (or about) different people, but this is not always so. In a sample study using interviews, respondents may say different things to different interviewers depending on the interviewer's sex, age, race, or other characteristics that act as extraneous variables in the research. Sometimes it is possible to avoid this problem by using a single interviewer or interviewers who are similar in terms of characteristics that may influence a person's response. Still, *holding the interviewer constant* is not always an ideal solution. Consider a survey on interracial attitudes. People will respond differently to black and white interviewers, but it would not help to use only one race of interviewers. Any difference between the responses of whites and blacks may be due either to their different attitudes or to their different reactions to being interviewed by, for example, a black (extraneous variable). In this case, control may be achieved either by eliminating the interviewer entirely (a mailed questionnaire could be used if its contents did not reveal its author's race), or by using both black and white interviewers to collect data from both black and white respondents. The latter solution may be preferable because it allows one to both hold the interviewer's race constant and to measure its effects. If black and white interviewers get the same results from similar respondents, it can be concluded that the interviewer's race made no difference. This method of using both races of interviewer has created an experiment within the sample study: Race of interviewer is manipulated to see its effects on respondents. The strategy of measuring the effect of a potential extraneous variable is discussed further on pages 93 to 94.

Other problems exist in trying to get the same information from different people. It may go without saying that in a questionnaire everyone should be asked the same

questions, but one cannot always assume that the questions are understood the same way by everyone. Some people may not understand because of limited vocabulary or reading ability, and it is not safe to assume their answers mean the same as those of other people. This problem can best be prevented by preliminary work by the researcher to make sure questions are understood.

The problem of "getting the same information" exists not only in interview and questionnaire situations, but in most forms of quantitative research, including laboratory experiments. In conducting an experiment on learning, for example, sounds from outside the lab may distract some subjects and constitute an extraneous variable that should be controlled because people are not all learning under the same conditions. You could do nothing, assuming that each subject is equally likely to be subjected to a distracting level of noise, or you could hold noise constant by placing the subject in a soundproof room, or putting plugs in his/her ears, or giving her/him a head set with a prerecorded tape of noise to listen to. When it is fairly easy to hold an extraneous variable constant by use of a standard procedure, this is the best method of control.

METHODS OF CONTROL

The information collected sometimes depends on who gathers it, or on how, when, or where it was collected. There are two main methods to control this problem.

Hold procedures constant. Ask questions in the same order, with the same wording; use one interviewer; always collect data in the same lab, at the same time of day, on the same apparatus, and so on.

Experimentally manipulate the variable causing responses to vary. The use of black and white interviewers in a survey of racial attitudes is the example used above. This method allows one to both control for and measure the effect of a potential extraneous variable.

When none of these controls is used, the researcher must assume (a better word is hope) that variations are random. The example of noise outside the learning lab is an illustration. With luck, this will influence the subjects in each group about equally. Even with luck, though, noise may so increase the variability between individuals as to hide any effect of the variables being studied.

The major validity problems of sample studies also exist in correlational and experimental research. All these quantitative methods operationalize variables, draw conclusions from samples to populations, and attempt to get the same information repeatedly. When evaluating quantitative research, watch out for:

Invalid operational definitions: ones that measure a variable other than or in addition to what they are supposed to measure.

Biased samples: samples that are systematically different in some way from the population they are drawn from.

Uncontrolled variation in information: information that depends on who collected it, or how, when or where it was collected.

If you suspect any of these problems, name the extraneous variable(s) that might be responsible, and try to explain how the extraneous variable(s) might account for the researcher's findings.

CORRELATIONAL RESEARCH: THE PROBLEM OF SUBJECT VARIABLES

Correlational research assesses the relationship between variables without manipulating any variable. The essential problem with this procedure is that it is impossible to measure one variable at a time in existing populations. Any measure of occupational status, for example, is in part a measure of education, because in our society the status of an occupation is closely related to the amount of education required for it. To attempt to relate occupational status to any other variable (say, intelligence or leadership ability) is difficult because any relationship that appears to exist may, in fact, be due to either occupation or education. Such variables as occupational status, education, intelligence, political party affiliation, and others, when they are measured as things a subject possesses before the research begins, are called *subject variables,* or *organismic variables,* and they pose the validity problem most characteristic of correlational research.

A *subject variable* or *organismic variable* is any characteristic that a research subject brings along to the research setting. For individuals, these characteristics include such attributes as sex, religion, education, and so on; for groups, they include group structure, communication patterns, and coalitions within the group. Some variables may or may not be organismic, depending on how they are treated in research. Anxiety is a good example. Consider an investigation of the effect of anxiety on learning. One method for this investigation would categorize people as highly anxious, moderately anxious, or nonanxious, using a pretest instrument such as the Taylor Manifest Anxiety Scale. The subjects would be given standard material to learn, and their performances compared. This correlational study measures the anxiety subjects bring with them to the study; anxiety is a subject variable. An experimental study of the effect of anxiety on learning might attempt to create anxiety experimentally (e.g., by misleading some subjects to believe that they are about to take an intelligence test), and measure the performance of anxious and control subjects on the same learning task. In such a study, anxiety is manipulated and is not treated as an organismic variable. Both studies use between-subjects *designs,* because they draw conclusions by comparing one group of people to another. Only the second study is a between-subjects *experiment* because only this study manipulates anxiety and assigns people to the anxious or nonanxious groups.

The variables in correlational studies tend to be organismic variables. The problem this creates is that other organismic variables are invariably correlated with those measured. Anxiety may be related to low self-esteem, insecurity in the presence of authority figures, emotional instability, or any number of other things. Therefore, if a correlational study shows anxiety to be related to learning, several alternative explanations are plausible. Learning may be affected by anxiety, by any of the correlated variables mentioned, by some other correlated variable, or by any combination of the above. *Measures of organismic variables are always confounded by other organismic variables. Whenever one variable in a hypothesis is organismic, all correlates of this variable are extraneous variables in the study, and each one can potentially be used to suggest an alternative explanation.*

METHODS OF CONTROL

The example of anxiety and learning suggests one possible way to eliminate the problem.

Use an experimental design with random assignment. When a subject variable is the independent variable in a hypothesis and it is capable of manipulation, a *between-subjects experiment* that employs random assignment (equivalent-groups design) can minimize the threat to validity by manipulating the subject variable. Thus, in an experiment on the effect of anxiety on learning, the potential subjects are randomly assigned to two groups: the experimental (anxious group) and the control (not anxious group). The two groups are treated identically except for the procedure used to create anxiety. This means that subjects are greeted the same way and are given instructions that differ only in one respect. Let's say the "anxiety group" is told that what they are about to do is an intelligence test, while the controls are told that the experimenter wants to compare two word lists to see if they are of equal difficulty (or some other presumably non-anxiety-arousing instruction). Both groups are given the same word lists to memorize, and their learning is tested in the same way. Thus, the only systematic difference between the two groups is in the part of the instructions that was intended to produce (or not to produce) anxiety.

What about organismic variables? These subjects, like the subjects in a correlational study, differ in the anxiety they had when they arrived, and also in self-esteem, relationships to authority, and every other variable that may be related to anxiety or learning. However, the subjects who began with high anxiety, say, were equally likely to be assigned to either group; so also were the subjects with low self-esteem, emotional instability, and so on. Thus, it would be unreasonable to conclude that the reason the people in the anxious group performed better was that they were trying to bolster their low self-esteem by performing well. There is no reason to believe that this group had lower (or higher) self-esteem than the other group. When a researcher assigns subjects at random to groups in a between-subjects experiment, we say that organismic variables (self-esteem, anxiety before the experiment began, etc.) are *randomized. Random assignment to conditions eliminates any bias that might systematically put similar people in the same group.* Contrast the randomized experiment with a correlational study of the same variables. In the correlational study, the highly anxious people probably have other personality characteristics in common as well (some possibilities have already been mentioned), and any of these could explain any observed difference in learning. In the experiment, personality characteristics are randomized. They are not systematically related to anxiety, because each personality type is equally likely to be in the anxious and nonanxious groups of the experiment. Thus, personality differences between groups are unlikely explanations of any differences in learning.

It is important to note that randomization does not eliminate all the personality differences between the two groups, but only ensures that each personality type or characteristic is *equally likely* to be in either group. On occasion, the people with low self-esteem, for example, will be put in the same group by the luck of the draw, but this does not happen systematically. The errors due to randomization in experiments resemble those produced by representative or random sampling in sample studies. Random error exists in both, but it is tolerable because it is not biased in either direction, and because its magnitude can be estimated with statistical techniques. In short, while randomization does not eliminate extraneous variables, and it does not keep them from varying, it does minimize their power to offer alternative explanations for research findings.

Randomization controls organismic variables only when they are independent variables, and when they can be manipulated. In many cases, these criteria are not met. For example, most of the variables of interest to sociologists and political scientists are either impossible or very difficult to manipulate. Think of doing an experiment to measure the effects of religion, social class, stigmatization, alienation, cultural conflict, or social disorganization. All are variables that subjects (people or societies) carry with them, and which pretty much must be studied as they are. A general strategy for controlling the effects of correlates of organismic variables that cannot be manipulated is matching.

Matching. This is the strategy of comparing individuals or groups who are equal in terms of an extraneous variable in order to rule this variable out as an explanation of a hypothesized relationship. Wrightsman (1969) used matching in a study done to discover whether supporters of George Wallace for President in 1968 upheld "law and order" as much in their daily lives as their candidate did in his campaign. In Nashville, the local government had passed an ordinance requiring all cars to display a tax sticker (cost: $15) beginning November 1, 1968, a few days before the election. Wrightsman's study was simple: he and his students went around to parking lots after the law went into effect and noted the presence or absence of tax stickers on cars with political bumper stickers. Wallace supporters (operational definition: Wallace sticker on car) obeyed the law significantly less frequently than Humphrey or Nixon supporters, or cars without bumper stickers. This is a correlational study. Neither variable (candidate supported, obedience to law) was manipulated. It follows that organismic variables entered the study with the subjects (cars), and that some of these may be related to the variables being studied. One such variable is socioeconomic status. Wrightsman reasoned that Wallace supporters in Tennessee tended to come from the working class, and they might, therefore, be less likely to have the $15 for the sticker. If this were true, the findings could be explained without reference to the Wallace supporters' lawfulness.

Wrightsman used matching to rule out this explanation. Wrightsman's observers were instructed to proceed by looking for a car in a parking lot with a political bumper sticker, recording the necessary information about the car, and then recording the same information about the car parked closest on its left that had no bumper sticker. It was reasoned that cars parked next to each other in the same lot would likely belong to people of similar socioeconomic status who were on similar errands. Thus, each car was matched with a single other car of presumably equal socioeconomic status. Wallace cars were less law-abiding than the cars parked on their left, while Nixon and Humphrey cars were more law-abiding than the cars parked on their left.

It is important to realize that matching controls only those variables that are matched. It may still be, for example, that Wallace supporters who used bumper stickers were more generally rebellious people than the average Wallaceite. Since Wallace was not a major party candidate, affixing a Wallace sticker may have taken a streak of rebelliousness. The Humphrey and Nixon sticker-users may have been

more typical of all supporters of their candidates. If this were true, Wrightsman's results would imply that it was rebelliousness that led some people both to use Wallace stickers and not to affix tax stickers. The conclusions would not apply to Wallace supporters in general. However far-fetched this hypothesis, the matching for socioeconomic status does nothing to rule out the alternative explanation based on rebelliousness.

Because matching controls only those variables that are matched, there is a practical limit to how many organismic variables can be controlled by matching. A researcher generally uses matching to control only those variables most likely to provide alternative explanations of the expected results. This is sometimes done even in experimental research, when a variable is so important that the researcher is unwilling to rely on randomization to equalize it. Such a situation might exist in research on learning, where intelligence is so important an organismic variable that subjects may be matched on it before being randomly assigned to experimental groups.

Occasionally, in an experimental study, a special sort of matching called the *yoked control* is used. Subjects are paired and then undergo treatments that are identical except for the independent variable. A good example is Brady's work with the "executive monkeys" (Brady, 1958). Two monkeys, strapped into identical apparatus, were either shocked or not shocked together. Although each monkey had a lever in front of it, only one lever had the power to turn off or prevent the shocks. (The monkeys with this lever—the "executives"—developed ulcers.) Another example comes from dream research. In studies in which subjects are deprived of dreaming, their sleep is also interrupted, so the two variables (dreaming and sleeping) are confounded. To control this, a second subject may be yoked with the dream-deprived subject so that whenever one subject starts to dream, both are awakened, regardless of whether the second subject is dreaming. Thus, both subjects are interrupted in sleep to the same extent, but only one is systematically dream-deprived.

The ultimate in matching, of course, is to *compare subjects with themselves*. This is possible in correlational research on such topics as emotional mood, intellectual development, and social change, all of which imply change over time within a single individual or society. Within-subjects experiments also control for subject variables by comparing subjects with themselves. More will be said below about this method for controlling organismic variables.

Statistical control of correlates of organismic variables. Similar in intent to matching are a number of procedures that attempt to accomplish matching after the fact. The researcher collects information about possible extraneous variables and then compares subjects who are equivalent in terms of these variables. Wrightsman used an elementary form of statistical control in his bumper-sticker study to deal with the extraneous variable of socioeconomic status. On the assumption that the age of a car was a good index of the socioeconomic status of its owner, Wrightsman recorded the model years of all cars observed. When Wallace supporters with new (less than four year old) cars were compared with Humphrey and Nixon supporters with new cars, the Wallaceites were less obedient of the law. The same relationship held when people with older cars were compared. By comparing groups of cars of the same age, Wrightsman was able to judge the relationship between political preference and obedience with socioeconomic status held constant. Since the relationship still held, one alternative explanation was ruled out.

In this procedure, Wrightsman did not match individual cars, but controlled status effects through data analysis. Status was measured (by age of car), and the data analysis was broken down according to status in the hope that, with status held constant, the hypothesized relationship would still hold.

More sophisticated methods of statistical control have been developed to deal with the problem of subject variables, and you will find them in reports of correlational research. The statistical procedures and rationales for such techniques as partial correlation (e.g., Friedman, 1972; Hays, 1963) and analysis of covariance (e.g., Kerlinger, 1973; Winer, 1962) are described in various books on research methodology, including those cited here.

Inclusion of extraneous variable(s) in the hypothesis. Randomization, matching, and the statistical controls discussed above all attempt to keep extraneous variables out of consideration. It is also possible to measure an extraneous variable specifically to assess its effect on the variables of the original hypothesis. Consider this example. An educator wished to study the effect of programmed instruction on performance in a college introductory psychology course. Since the researcher did not have the power to see that students were randomly assigned to programmed or nonprogrammed instruction, the study was correlational. The final exam performance in a section receiving programmed instruction was compared with that of another section receiving more traditional instruction. The programmed group scored higher on the final exam. It was later discovered that the students in the programmed section had higher verbal ability (as measured by Scholastic Aptitude Tests). Thus, their success on the final exam might have been due either to superior instruction or to superior verbal ability; the two variables were confounded.

Because the researcher had information on all three variables (type of instruction, verbal ability, and performance) for each subject, it was possible to examine the joint effect of the independent variable and the "third" variable on performance. This was done by dividing the students in each section into subgroups according to their SAT Verbal scores, and by summarizing the results (see Table 3.4).

This table summarizes what we already know and gives additional information. The last column shows that the programmed instruction group scored higher on the final exam (80.8 to 73.1), and the bottom row shows that students with high verbal ability (SAT Verbal over 500) did better than students of lower verbal ability (83.3

Table 3.4. Mean Final Examination Scores of Introductory Psychology Students of Low and High Verbal Ability Receiving Two Types of Instruction (Hypothetical Data)

Method of Instruction	SAT Verbal Scores: 500 or Below	n^a	Over 500	n^a	Grand Mean
Programmed	77.0	9	83.0	16	80.8
Traditional	68.0	17	84.0	8	73.1
All students	71.1	26	83.3	24	76.6

[a] "n" denotes the number of people in each subgroup. In the programmed instruction section, nine students had SAT Verbal scores of 500 or below, and sixteen had scores over 500, and so on.

to 71.1). We can also see from the columns labeled *"n"* that more high-verbal students were in the programmed section (16 to 8, though each section had 25 students). This is information we already had. The special value of this table is that it also gives quantitative information on the effects of programmed instruction on each type of student (low and high verbal ability) separately. When we examine these data, we find that programmed instruction greatly improved the performance of low-verbal students, who scored 77, compared to 68 for similar students in the traditional class. However, programmed instruction was no help to the students with high verbal ability. In fact, these students performed slightly better in the traditional class (84 to 83). By measuring verbal ability, and including it as a variable for study, we have discovered that the effect of programmed instruction depends on the type of student being taught. The researcher started with a simple question about two variables—"Which instructional method is more effective?"—and was able to get an answer about three variables—"Programmed instruction is better with students of low verbal ability, but the method of instruction makes little difference with students of higher verbal ability." The joint effect of verbal ability and instructional method on performance is called an *interaction of variables*.

When the effect of one variable depends on the presence, absence, or amount of another variable, the two variables are said to interact. In the example, the effect of programmed instruction depends on the type of student. The reverse is also true: the performance of a given type of student depends on the type of instruction (at least for low-verbal students). An interaction exists whenever two or more independent variables, by virtue of acting at the same time, influence a dependent variable. In the programmed instruction example, both type of instruction and verbal ability are considered as independent variables which, when combined, have an influence on performance. That is, the effect of the two variables acting together is different from the sum of two separate effects. Programmed instruction increases performance, and so does verbal ability, but programmed instruction, when combined with high verbal ability, does nothing to increase performance.

Probably the most famous interaction is that of alcohol and barbiturates. Someone who is used to taking either drug knows what to expect when taking one alone, but the deaths of people who have taken both are proof that the interaction is different from the sum of the two drug effects. Each drug has an effect, and so does their interaction.

In the case of the drug interaction, each drug has an effect by itself, but the effect of the two taken together cannot be predicted from the effect of the single drugs alone. It is also possible for variables to interact even when it seems that neither of them has any effect by itself. Consider the example of a psychologist who tried out a new "energizing" drug on a sample of emotionally disturbed and mentally retarded children. The children's behavior changed after ingesting the drug, but the average change was zero. When the researcher divided the results between boys and girls, it became clear that the girls all became more active after taking the drug, while the boys became less active. When boys and girls were considered together, the increase and decrease canceled each other, and the net effect appeared to be zero. (This example and several others appear in a detailed article on the concept of interaction by Schaefer [1976].)

It is important to realize that the existence of an interaction changes the meaning

of the information that was available before the interaction was examined. This is obvious in the example of the "energizing" drug, where the effect of the drug seems to be zero until the interaction with the child's sex is taken into account. The point applies equally well to the example of programmed instruction. Although it is true that, for the students studied, those receiving programmed instruction did better, this simple statement is misleading. The facile interpretation—that programmed instruction helps students learn better—is incorrect. If there is any causal relationship, it can only be for some (low-verbal) students. By explicitly studying the extraneous variable of verbal ability, something was learned about programmed instruction that would have been missed if verbal ability had been controlled by randomization or matching. The strategy of including additional variables in the hypothesis sometimes allows us to discover that the effect of an independent variable may depend on a variable that had previously been thought to be extraneous. Since extraneous variables commonly interact in this way with variables of more direct interest, the best way to handle "extraneous" variables is often to explicitly measure them to assess their importance. This strategy reaches its highest development when an extraneous variable can be experimentally manipulated to study its effects, as in the example of black and white interviewers on page 87.

A common method of including extraneous variables in the hypothesis is *multiple regression analysis*. It is beyond the scope of this book to explain the mathematics of this common technique of economics, sociology, and political science, but we can briefly explain how it works in general terms. An example will help. In a study we conducted with our colleague Tom Dietz (Stern, Dietz, and Kalof, 1993), we surveyed a sample of college students to examine the how different beliefs about the consequences of changes in the natural environment affected the students' willingness to take action on environmental problems. Our chief measure of action was a scale of political behavior made up of the respondents' expressed willingness to participate in proenvironmental demonstrations, contribute money, sign petitions, and take a job with a company that harms the environment. (We reversed the responses on the last survey item so we could combine them by addition.) We measured three kinds of beliefs as independent variables: beliefs that environmental problems cause harm to the respondent or his or her family (showing an egoistic concern), beliefs that other, more distant human beings might be harmed (altruistic concern), and beliefs that environmental conditions harm nonhuman organisms or the biosphere generally (biospheric concern). We reasoned that students would be more likely to take action when they believed an environmental problem threatened things they valued. Because some of the literature on environmentalism claims that women are more proenvironmental than men, we recorded the gender of each respondent. At first, we intended to treat gender as an extraneous variable, but we arranged our analysis so that we could interpret it as an independent variable as well. We thus had one dependent variable (willingness to act politically) and four variables treated as independent variables (three kinds of concerns or values, and gender).

We wanted to learn whether each kind of value had an effect on proenvironmental political behavior independently of the effects of the other values and independently of any effect of gender. (Actually, since this is a correlational study, we were measuring associations and not effects, but our hypothesis was causal and we intended to interpret any associations as effects.) We could have presented the data in a table

like Table 3.4, but with four independent variables, it would be difficult to make sense of the numbers in the table. Instead, we used multiple linear regression analysis. This is a statistical technique that arrives at numbers (called *regression coefficients*) for each independent variable that allow a researcher to estimate each respondent's score on the dependent variable by multiplying the respondent's score on each independent variable by its coefficient and then adding. In our study, multiple regression results in something like this:

$$\begin{aligned}&(e \times \text{egoistic concern score})\\ +\ &(s \times \text{social-altruistic concern score})\\ +\ &(b \times \text{biospheric concern score})\\ +\ &(g \times \text{gender [0 if male, 1 if female]})\\ =\ &\text{estimated political action score.}\end{aligned}$$

In this equation, e, s, b, and g are the regression coefficients for each of the four independent variables. The equation allows the researcher to estimate an individual's political action score given knowledge of that individual's beliefs and gender. Larger coefficients mean that an independent variable has a large effect; coefficients close to zero mean that it has little effect on the political action score. In terms of controlling for extraneous variables, the importance of this equation is that each regression coefficient takes the other coefficients into account—it represents the effect of a particular independent variable with the other ones held constant by a statistical procedure.

In our study, we first compared women and men on the political action scale, and found that the women students expressed more willingness than the men to take proenvironmental action. The political action scale was a standardized scale, which means that the average score is zero, and 95% of the individuals have scores between −2 and +2. On average, women scored +.23 and men −.28, a difference of .51, which is strongly statistically significant. We then conducted the multiple regression with the three kinds of beliefs and gender as independent variables. It showed that each kind of belief had a statistically significant association with political behavior, but that gender had no significant effect. With beliefs controlled statistically, the average score for women was +.07 and for men, −.10, a difference of only .17.

We concluded several things from this. First, gender has some sort of effect on proenvironmental political behavior. Second, beliefs about all three kinds of effects of environmental conditions—on self, on others, and on the biosphere—also affect political behavior. Third, when gender and beliefs are considered together, the effect of gender disappears. We guessed that gender had an *indirect effect* on proenvironmental behavior by influencing environmentally relevant beliefs: that is, gender directly affected environmental beliefs, but only beliefs (and not gender) affected the behavioral measure directly. In support of this guess, we found that the college women in this population believed that environmental problems had more serious effects on self, on others, *and* on the biosphere than the men did.

This example shows how multiple regression techniques can be used to consider the effect of an extraneous variable (gender) by including it in the hypothesis. It also shows how regression can lead to a deeper understanding than can be obtained by holding the variable constant (for example, studying only female students), or by using techniques like matching that achieve control of the variable but not information about it. We concluded that gender was *not* extraneous to environmental concern. It seemed to somehow influence beliefs about the environment that, in turn, affected behavior. This conclusion has led us to conduct more research to try to understand the relationships that might explain the role of gender in environmental attitudes and behavior. In our new research, we are trying to learn more about differences between men's and women's beliefs about the environment.

The kind of findings we observed are sometimes reported in research articles as a *path diagram,* which represents statistically significant associations in a correlational study that are presumed to reflect causal relations with arrows, and the strengths of those associations with regression coefficients. Figure 3.1 is a path diagram that summarizes findings from our study. Each of the three arrows on the left side of the diagram represents a regression equation estimating scores on a belief scale from gender. The three arrows on the right side of the diagram represent the single regression equation estimating political action scores from the three kinds of beliefs *and* from gender. There is no arrow directly from gender to the political action scale because that regression coefficient was not statistically significant. So the diagram graphically represents our interpretation that gender has an effect on commitment to action indirectly, through environmental beliefs, but not directly, independent of beliefs.

The strategy of including extraneous variables in the hypothesis is limited in that it controls only those variables that are included. Other organismic variables are left uncontrolled. In the programmed instruction study, for example, an important organismic variable is the teacher's instructional style. Type of instruction is confounded with the style and personality of the teachers involved, and this cannot be changed by measuring students' verbal ability. In the study of environmentalism, political liberalism or conservatism is an extraneous variable that is left uncontrolled. It may affect both environmental beliefs and behavior, and it is not controlled by including gender in the hypothesis.

Often, researchers begin with a hypothesis stated in terms of the interaction of variables. The critical period hypothesis in developmental psychology is an example.

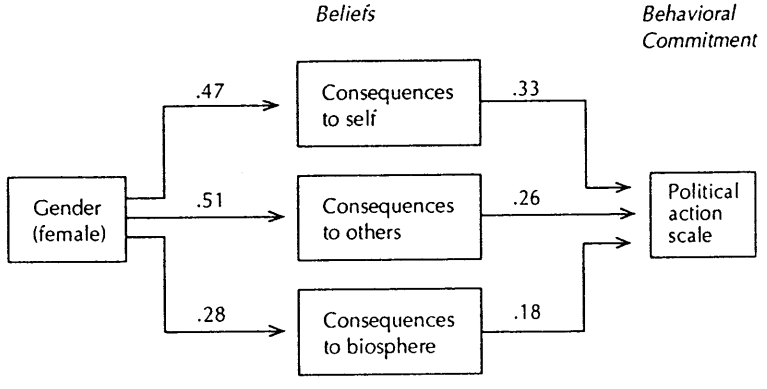

Figure 3.1 Effects of gender and beliefs on commitment to take proenvironmental action. (All the regression coefficients reported are statistically significant at the .05 level or beyond.

The effect of a life experience is held to be dependent on the stage of development during which it occurs. That is, the effect of experience depends on time. There are many other examples in the social sciences in which interactions are hypothesized. Such hypotheses can be tested either by correlational research, as in the programmed instruction example, or by experimental methods, if the independent variables can be manipulated. In any research of this type, the presence of an interaction changes the meaning of any effects of the variables that interact. The reasoning behind this is the same whether the interaction was predicted or not.

In summary, the chief problem in drawing conclusions from correlational research is that such research measures organismic variables. Any finding explained in terms of an organismic variable may alternatively be explained in terms of any other organismic variable that is correlated with the first, but was not studied. Four strategies are commonly used to solve this problem of inference:

Randomization. A between-subjects experiment allows organismic variables to vary randomly, and eliminates systematic error.

Matching. Individuals can be matched with others (or themselves) so that the only important differences are in terms of variables in the hypothesis.

Statistical control. Individuals can be matched after the fact to compare the effect of an independent variable on people who were initially comparable in terms of selected extraneous variable(s).

Inclusion of extraneous variable(s) in the hypothesis. An extraneous variable is measured and treated as an independent variable that may, either alone or in interaction with other variables in the study, influence a dependent variable of interest.

WITHIN-SUBJECTS EXPERIMENT: THE PROBLEM OF "TIME-TIED" EXTRANEOUS VARIABLES

The within-subjects experiment has already been mentioned as a technique for controlling organismic variables. By observing changes in a single individual, or group of individuals, the effects of a manipulated variable can be measured while achieving perfect control of organismic variables. A price is paid for this control, however. When a subject is observed over a period of time, to see when he/she changes, any variable that might have produced an observed change is confounded with the passage of time. The independent variable might have changed the subject, but any other events during the same time period may also be responsible. In short, *"time-tied" extraneous variables* (Agnew & Pyke, 1969) *pose the characteristic validity problem of within-subjects experiments.*

Consider this example of a within-subjects experiment. A researcher is studying the effect of a new drug, Memoraid, on learning. Because some people learn faster than others, the researcher decides to use subjects as their own controls. Each subject will get a chance to learn both with and without the drug. On the first day, each subject is tested with a placebo (no drug), on a task involving the learning of a series of nonsense syllables. On the next day the subjects are given the real drug, and are asked to learn a new list of nonsense syllables. The subjects learn better the second time. The experimenter might be tempted to conclude that Memoraid improves learning. However, many time-tied extraneous variables are confounded with the drug's effect. Subjects may have learned something about memorizing nonsense syllables on the first day and applied the knowledge by learning better on the second day. Or, they may have become bored with nonsense syllables. This would mean that the drug has a stronger effect than the results indicate. Subjects may have gotten to know the experimenter better, and, feeling comfortable, may have performed better the second time. On the other hand, familiarity may have lowered their anxiety level, leaving them less motivated to perform well. It's also possible that time-tied changes took place in the apparatus used to collect data. The wires in the memory drum used to display the nonsense syllables may have become worn, and resistance in the circuits might have increased, causing the drum to move slower and giving subjects more time to learn on the second day. The laboratory might have been visited by noisy plumbers on the first day, or the weather might have been rainy, making the subjects mentally sluggish. And so on and so on.

The above problems are not unique to experimental research; they also exist in case studies, naturalistic observations, and correlational research. Suppose the effects of Memoraid were first discovered by a scientist who accidentally ingested some of the drug. Her/his evidence would have been based on a retrospective case study, without *any* effort at systematic control. The effects attributed to the drug could have easily been due to any number of time-tied variables.

METHODS OF CONTROL

Three common procedures for controlling time-tied extraneous variables are described below.

Use of a comparison group. A group of subjects could be observed over the same time period as the experimental subjects, but without exposure to the independent variable. In the Memoraid experiment, this could be accomplished by randomly assigning subjects to get either the drug *or* the placebo (control). This would transform the study into a between-subjects experiment. The between-subjects design does not have serious problems with time-tied extraneous variables.

Comparison groups can be used to control time-tied extraneous variables in nonexperimental research, and some special designs have been developed for this purpose. In the *multiple time-series design* (Campbell & Stanley, 1963; Gottman, McFall, & Barnett, 1969), a group that has been exposed to an independent variable is compared to a control group on several occasions both before and after exposure. Both groups are exposed to the passage of time, but if the independent variable makes a difference, they will change differently over time, and especially after the independent variable is introduced. More sophisticated procedures for making unambiguous inferences from correlational data over time include the cross-lagged panel design and methods of path analysis (e.g., Heise, 1969; Land, 1969).

Counterbalancing. In a counterbalanced design, two or more groups are used, one for each treatment condition. However, unlike the between-subjects experiment in which each group is exposed to different treatments, in a counterbalanced design, each group gets all treatments, but in different orders. To counterbalance the experiment on Memoraid and learning, one group would get the placebo on the first day, and Memoraid on the next day. The other subjects would get Memoraid first, and

then the placebo. Thus, any effect of learning to memorize, or boredom, or noisy plumbers, or slow machines would be equally divided between subjects getting the drug and subjects getting the placebo. If any of these variables either aids or interferes with learning, it could not explain any difference between drug and placebo treatments. It is also possible to counterbalance the lists of nonsense syllables in this study. After all, one list may be easier to learn than the other. To control for this possibility, the two groups can be divided, with half of each group learning list A first, and then list B. The order would be reversed for the other half of each group. The counterbalanced design for the Memoraid experiment is given in Table 3.5.

ABA design (repeated experiments). In an ABA design, subjects are observed before and after an experimental treatment, as well as while they are getting the treatment. (A represents the condition in which the treatment is absent and B the condition in which it is present.) This experimental design is common to most studies in the field of behavior modification. For example, suppose a teacher wants to decrease the frequency of aggressive outbursts by one of the boys in the class. The teacher plans to reinforce nonaggressive behavior with praise, and remove the boy from his classmates when he acts aggressively toward them. First, a "baseline" is taken. That is, the child is observed for a while before the treatment begins, and the frequency of aggressive outbursts is tabulated. When treatment begins, the number of aggressive acts each day is recorded, and, it is hoped, it decreases. The experiment may go on, alternately starting and stopping the treatment a few times, to demonstrate a consistent relationship between onset of treatment and decreases in aggressive behavior. The subject has served as his own control.

Note that in this example, the within-subjects design is used with only one subject. Partly because of increased interest in behavior modification techniques, psychologists have paid considerable attention to experimental designs for single subjects. As a result, they have developed some subtle methods of within-subject experimentation (e.g., Kratochwill & Levin, 1992).

The ABA design controls for some of the important time-tied variables. If a time-tied variable operated continually, its effect should increase with time, independent of treatments. In the Memoraid experiment, the effect of learning to learn, or boredom, or acquaintance of the subject and the experimenter should get stronger and stronger, rather than coming and going with the drug. If subjects were given a placebo at both ends of the experiment, and if they learned best in the middle (under the drug), the above-mentioned variables could probably be discounted.

Both counterbalancing and the ABA design require that variables be manipulated. That is, these controls can be used only in experimental research. In correlational research, case studies, and naturalistic observations, the comparison group strategy

Table 3.5. A Counterbalanced Design

	First-Day Treatment		Second-Day Treatment	
Group	Drug	Syllable List	Drug	Syllable List
No. 1	Placebo	List A	Memoraid	List B
No. 2	Placebo	List B	Memoraid	List A
No. 3	Memoraid	List A	Placebo	List B
No. 4	Memoraid	List B	Placebo	List A

is the only feasible way to rule out alternative explanations dependent on the passage of time.

BETWEEN-SUBJECTS EXPERIMENTS: THE IMPORTANCE OF GROUP EQUIVALENCE

Recall from Chapter 2 that unless a between-subjects experiment randomly assigns subjects to treatment and comparison groups (the equivalent groups design), it cannot be assumed that the groups are comparable. In a nonequivalent groups design, subjects assigned to different treatments may be systematically different in terms of whatever organismic variables are associated with the particular group they are in. But between-subjects experiments often use naturally occurring comparison groups. In such experiments, it is generally possible to think of alternative explanations that depend on the noncomparability of the groups (Cook & Campbell, 1979). A few examples:

In educational research, treatments are often applied to whole classrooms, with other classrooms providing the comparison group. But pupils are rarely assigned to classrooms at random. They may be assigned by ability grouping, by pressure from parents who try to get their children into the classroom of the teacher they think is best, or by some other nonrandom procedure. Every deviation from randomness introduces extraneous variables into the experiment. In these examples, pupil ability and parental tendencies toward intervention may influence the dependent variable in an educational experiment as much as the independent variable does.

In marketing research, companies sometimes test-market a new version of a product in one city and use another city as the comparison group. When this happens, every difference between the cities is a confounding variable in the experiment. Among the variables that might matter in marketing are average income and educational levels, unemployment levels, and, for some products, cultural or religious variables that affect purchases of the product.

Sometimes a government agency or a company evaluates a new program by comparing the first people who participate with others in their vicinity. For example, to evaluate an energy-conservation program, a natural-gas company offered it to all its customers in a city. It surveyed the first 200 participants in the program and a random sample of 200 other households to see how many energy-conservation activities they had taken in their homes, and it attributed the difference between the groups to the program. But participation in the program was not by random assignment, and there are systematic differences between the participants and the nonparticipants. A very important difference is that the participants were obviously among the households most interested in energy conservation. Some of the energy conservation measures these households took might well be due to their own interest in conservation, and not to the program.

METHODS OF CONTROL

Matching. When random assignment is not possible or desirable in between-subjects experiments, the most common method of control for organismic variables is matching. As discussed in methods of control for correlational studies, matching subjects

on important extraneous variables that might affect the dependent variable eliminates alternative explanations based on those extraneous variables. Using only first-year college students in the dormitory crowding experiment conducted by Baum and Valins (1977) eliminated the possibility that the student's year in college might have caused the observed differences between the crowded and uncrowded dorm residents. For example, third- or fourth-year dorm residents might feel less crowded than first-year residents because they have had more time to become adapted to dorm conditions.

Waiting-list control. Another method, useful in program-evaluation studies like the energy-conservation experiment above, is the waiting-list control group. The gas company might have randomly chosen half of its first 200 participants to be told that the program was oversubscribed and that they would have to wait a month to join. It could then have compared the energy activities these people undertook at the end of the month with those undertaken by the other half of the 200, who were allowed to join the program immediately. This kind of control achieves many of the benefits of random assignment to treatments. It automatically controls for interest in energy conservation, because all 200 customers can be assumed to be about equal on that variable and because the most highly interested were equally likely to be assigned to the treatment or the control group. But it is not as good a research design as one that randomly assigns treatments to a representative sample of all the company's customers, because the subjects in both groups are probably systematically different from the rest of the population that the company wants to learn about. What works well with the highly motivated first 200 might not work at all in the rest of the city.

Statistical control. Yet another way to eliminate alternative explanations in between-subjects, nonequivalent-groups experiments is through statistical control of organismic variables. Usually this is accomplished by including the variables in the data analysis. For example, in their quasi-experiment on crowding in dorms, Baum and Valins (1977) used a survey to collect background information on the dorm residents. The researchers included this information in the data analysis, and none of the background variables had an effect on the dependent measures (Epstein & Baum, 1978). As with correlational studies, the most common methods of statistical control for quasi-experiments are cross-tabulation, partial correlation, regression, and analysis of variance. More sophisticated statistical analyses have recently been developed. Many of these techniques have been reviewed by Achen (1986).

Even when subjects are randomly assigned to experimental treatments, the between-subjects experiment is not a foolproof design. "On stage" and other changes caused by research, as well as biases in judgment, are problems for this method. Just as for other quantitative methods (sample study and correlational study), sampling bias and invalid operational definitions can pose serious threats to the validity of a between-subjects experiment, with or without randomization.

The last sections of this chapter have outlined the major sources of alternative explanations for the results of quantitative research. Table 3.6 summarizes these sources, indicates when they are most likely to cause trouble, and describes some of the methods researchers use to rule out alternative explanations. This table, together with Tables 3.1, 3.2, and 3.3, constitute a summary of the common sources of alternative explanations for the findings of social scientific research.

The exercises for this chapter emphasize the ability to identify alternative explanations for research results and to suggest controls that would improve the research. Specifically, the questions in the exercises directly test your ability to use the following terms:

Table 3.6. Sources of Alternative Explanation in Quantitative Research Methods

Source	Description	When a Problem	Methods of Control
Invalid operational definition	Operational definition measures another variable as well	Potential problem with all operational definitions	Prevent confounding
Sampling bias	Sample systematically different in some respects from population	Nonrandom sampling; volunteer subjects	Assure that sample is representative
Uncontrolled variation in information	Results depend on who collected them or how, when, or where collected	Whenever data are collected on several people or occasions	Hold procedures constant; manipulate the variable responsible
Organismic variables	Subject variables are measured and confounded with extraneous variables	When hypothesis contains an organismic variable; nonequivalent groups experiments	Randomization; matching; subject as own control; statistical control; waiting-list control; manipulate extraneous variable
Time-tied variables	Independent variable's effects confounded with the passage of time	Research with no comparison group	Add comparison group; counterbalancing; ABA design

Extraneous variable (defined on p. 63)
Alternative explanation (p. 63)
Holding procedures constant (p. 88)
Matching (p. 91)
Statistical control (p. 92)
Subjects as their own controls (p. 92)
Comparison group as a control (p. 99)
Sampling bias (p. 86)

The main point of this chapter is to build your skill in reading scientific reports with a critical eye to alternative explanations for reported findings. The exercises provide the opportunity to practice evaluating reports of research. A procedure for seeking alternative explanations may be helpful to you, until you develop enough experience to look in the right places.

Begin by identifying the research method used in the study you are evaluating. Table 3.7 identifies the most common sources of alternative explanation for each research method, and suggests the place to begin looking for validity problems in a

Table 3.7. Sources of Alternative Explanation in Six Methods of Scientific Research

Source	Research Method						
	Naturalistic Observation	Retrospective Case Study	Sample Study	Correlational Study	Within-Subjects Experiment	Between-Subjects, Nonequivalent Groups	Between-Subjects, Equivalent Groups
On stage effects (p. 65)	*	*	*	*	*	*	*
More persistent changes due to research (p. 69)	*	*	*	*	*	*	*
Incomplete access (p. 77)	**	**	*	*	*	*	*
Researcher selectivity (p. 78)	*	**	—	—	—	—	—
Researcher distortion (p. 80)	*	**	*	*	*	*	*
Selective or distorted memory (p. 82)	A source of alternative explanation when research relies on memory data						
Invalid operational definitions (p. 84)	—	—	**	**	**	**	**
Sampling bias (p. 86)	—	—	***	*	*	*	*
Uncontrolled variation in information (p. 87)	—	—	*	*	*	*	*
Organismic variables (p. 89)	*	*	—	***	— (Held constant)	**	—
Time-tied Variables (p. 98)	*	*	—	* (If subjects compared to selves)	***	**	—

*Asterisks indicate common sources of alternative explanation in a particular research method. The more asterisks, the more serious the problem. In looking for alternative explanations, start with sources given two or three asterisks, but do not ignore any column with an asterisk.

piece of research. By using the information in Table 3.7, you should be able to ask questions that will lead you to alternative explanations. With practice, you should become able to ask these questions and think of alternative explanations without consulting the table.

EXERCISES

The objective of this chapter and the following one is to help you learn to critically evaluate reports of empirical research. The exercises in this chapter provide practice on prepared summaries of (usually imaginary) research. In Chapter 4, you will apply your skills to actual published scientific reports.

The central skill in evaluating research is the skill of identifying plausible alternative explanations for researchers' findings. Because this ability is of greatest importance, the "briefer" exercises and problems that follow are provided for practice in offering alternative explanations. These are followed by other exercises that use all the skills of this chapter.

BRIEFER EXERCISES

For each research report summarized below, answer these questions in the spaces provided:

(a) What method was used to collect the data? (Use categories defined in Chapter 2.)
(b) What is the hypothesis (if any)? Identify the variables in the hypothesis, labeling the independent and dependent variables if the hypothesis is causal.
(c) Identify the findings (that is, what relationship of variables was observed?).
(d) Identify extraneous variables that suggest alternative explanations of these findings. State the alternative explanations, and suggest one method to control for each extraneous variable.

1. To assess the effects of psychotherapy as opposed to drug therapy, the progress of schizophrenic patients receiving these therapies was observed. All subjects were diagnosed schizophrenic, and were assigned to the treatments considered appropriate by their attending physicians. Subjects in the drug therapy condition were receiving a variety of drug treatments, but none was receiving psychotherapy. Those in the psychotherapy treatment included only subjects for whom no drugs were prescribed. After six months' observation, improvement, as judged by outside consultant psychiatrists, was greater in the psychotherapy group. It was concluded that psychotherapy is more effective than drug therapy in the treatment of schizophrenia.

(a)