CHAPTER FOUR

Observation

CHAPTER OUTLINE

Overview

We observe behavior every day. Admit it. Many of us are people watchers. And it isn't simply because we are dedicated voyeurs or even exceptionally curious, although human behavior is certainly often interesting. People's behaviors—gestures, expressions, postures, choice of apparel—contain a lot of information as popular books on "body language" seek to emphasize (e.g., Pease & Pease, 2004). Whether it is a simple smile or a subtle courtship ritual, another person's behavior frequently provides cues that are quickly recognized. Indeed, research reveals that many of our expressions are "universal" signals, that is, recognized in all cultures (e.g., Ekman, 1994). Scientists, too, rely on their observations to learn a lot about behavior (although see Baumeister, Vohs, & Funder, 2007, for an opinion that psychologists don't observe actual behavior enough).

Our everyday observations and those of scientists differ in many ways. When we observe casually, we may not be aware of factors that bias our observations. Moreover, we rarely keep formal records of our observations. Instead, we rely on our memory of the events even though our own experience (and psychological research) confirms that our memory is not perfect!

Scientific observation is made under precisely defined conditions, in a systematic and objective manner, and with careful record keeping. The primary goal of observational methods is to describe behavior. Scientists strive to describe behavior *fully* and as *accurately* as possible. Researchers face serious challenges in reaching this goal. Clearly, it is impossible for researchers to observe *all* of a person's behavior. Scientists rely on observing *samples* of people's behavior, but they must decide whether their samples represent people's *usual* behavior. In this chapter we describe how scientists select samples of behavior. Researchers face a second challenge in trying to describe behavior fully: Behavior frequently changes depending on the situation or context in which the behavior occurs. Consider your own behavior. Do you behave the same at home as in school, or at a party compared to in a classroom? Does your observation of others, such as your friends, lead you to conclude that context is important? Have you observed that children sometimes change their behavior when they are with one or the other of their parents? Complete descriptions of behavior require that observations be made across many different situations and at different times. Observation provides a rich source of hypotheses about behavior, and so observation can also be a first step in discovering why we behave the way we do.

In this chapter you will see that the scientist-observer is not always passively recording behavior as it occurs. We will take a look at reasons why scientists intervene to create special situations for their observations. We'll also look at ways to investigate behavior that do not require direct observation of people. By examining physical traces (e.g., graffiti, textbook underlining) and archival records (e.g., marriage licenses, high school yearbooks), scientists gain important insights into people's behavior. We also introduce you to methods for recording and for analyzing observational data. Finally, we describe important challenges that can make it difficult to interpret the results of observational studies.

Sampling Behavior

- When a complete record of behavior cannot be obtained, researchers seek to obtain a representative sample of behavior.
- The extent to which observations may be generalized (external validity) depends on how behavior is sampled.

Before conducting an observational study, researchers must make a number of important decisions about when and where observations will be made. Because the investigator typically cannot observe all behavior, only certain behaviors occurring at particular times, in specific settings, and under particular conditions can be observed. In other words, behavior must be *sampled*. This sample is used to *represent* the larger population of all possible behaviors. By choosing times, settings, and conditions for their observations that are representative of a population of behaviors, researchers can *generalize* their findings to that population. That is, results can be generalized only to participants, times, settings, and conditions *similar* to those in the study in which the observations were made. The key feature of *representative samples* is that they are "like" the larger population from which they are drawn. For example, observations made of classroom behavior at the beginning of a school year may be representative of behavior early in the school year, but may not yield results that are typical of behavior seen at the end of the school year.

Key Concept

External validity refers to the extent to which the results of a research study can be generalized to different populations, settings, and conditions. Recall that validity concerns "truthfulness." When we seek to establish the external validity of a study, we examine the extent to which a study's findings may be used accurately to describe people, settings, and conditions beyond those used in the study. In this section we describe how time, event, and situation sampling are used to enhance the external validity of observational findings.

Time Sampling

- Time sampling refers to researchers choosing time intervals for making observations either systematically or randomly.
- When researchers are interested in events that happen infrequently, they rely on event sampling to sample behavior.

Key Concept

Researchers typically use a combination of time sampling and situation sampling to identify representative samples of behavior. In **time sampling,** researchers seek representative samples by choosing various time intervals for their observations. Intervals may be selected systematically (e.g., observing the first day of each week), randomly, or both. Consider how time sampling could be used to observe children's classroom behavior. If the researchers restricted their observations to certain times of the day (say, mornings only), they would not be able to generalize their findings to the rest of the school day. One approach to obtaining a representative sample is to schedule observation periods *systematically* throughout the school day. Observations might be made during

four 30-minute periods every 2 hours. A *random* time-sampling technique could be used in the same situation by distributing four 30-minute periods randomly over the course of the day. A different random schedule would be determined for each day observations are made. Times would vary from day to day, but, over the long run, behavior would be sampled equally from all times of the school day.

Electronic devices provide a major advantage in carrying out time sampling using randomization. Electronic pagers can be programmed to signal observers on a random time schedule (normal sleeping times are excluded). For example, in their study of middle-class youth, Larson and others (Larson, Richards, Moneta, Holmbeck, & Duckett, 1996) obtained self-reports on adolescents' experiences at "16,477 random moments" in their lives. Systematic and random time-sampling procedures are often combined, as when observation intervals are scheduled systematically but observations within an interval are made at random times. For example, electronic pagers might be programmed to signal every 3 hours (systematic), but at a randomly selected time during each 3-hour interval. Whatever time-sampling procedure is used, the goal of time sampling is to obtain a representative sample of behavior that will represent an organism's usual behavior.

Time sampling is not an effective method for sampling behavior when the event of interest occurs infrequently. Researchers who use time sampling for infrequent events may miss the event entirely. Or, if the event lasts a long time, time sampling may lead the researcher to miss an important portion of the event, such as its beginning or end. In *event sampling* the observer records each event that meets a predetermined definition. For example, researchers interested in observing children's reactions to special events in school, such as a holiday play, would use event sampling. The special event defines when the observations are to be made.

Event sampling also is useful for observing behavior during events that occur unpredictably, such as natural or technical disasters. Whenever possible, observers try to be present at those times when an event of interest occurs or is likely to occur. Although event sampling is an effective and efficient method for observing infrequent or unpredictable events, the use of event sampling can easily introduce biases into the record of behavior. For instance, event sampling could lead an observer to sample at the times that are most "convenient" or only when an event is certain to occur. The resulting sample of behavior at these times may not be representative of the same behavior at other times. There is yet another sampling procedure that also may be used to obtain a representative sample: situation sampling.

Situation Sampling

- Situation sampling involves studying behavior in different locations and under different circumstances and conditions.
- Situation sampling enhances the external validity of findings.
- Within situations, subject sampling may be used to observe some people in the setting.

Key Concept

Researchers can significantly increase the external validity of observational findings by using situation sampling. **Situation sampling** involves observing behavior in as many different locations and under as many different circumstances and conditions as possible. By sampling various situations, researchers reduce the chance that their results will be unique to specific circumstances or conditions. For example, animals do not behave the same way in zoos as they do in the wild or, it seems, in different locales. This is seen in studies of mutual eye gaze between mother and infant chimpanzees. Mutual eye gaze occurs in chimps as it does in humans, but in one study of chimpanzees the frequency of this behavior differed between animals observed in the United States and in Japan (Bard et al., 2005). Similarly, we can expect human behavior to differ across different settings.

By sampling different situations, a researcher can also increase the diversity of the subject sample and, hence, achieve greater generality of findings than could be claimed if only particular types of individuals were observed. For example, LaFrance and Mayo (1976) investigated racial differences in eye contact and sampled many different situations. Pairs of individuals were observed in college cafeterias, business-district fast-food outlets, hospital and airport waiting rooms, and restaurants. By using situation sampling, the investigators were able to include in their sample people who differed in age, socioeconomic status, sex, and race. Their observations of cultural differences in eye contact have considerably greater external validity than if they had studied only certain types of participants in only a specific situation.

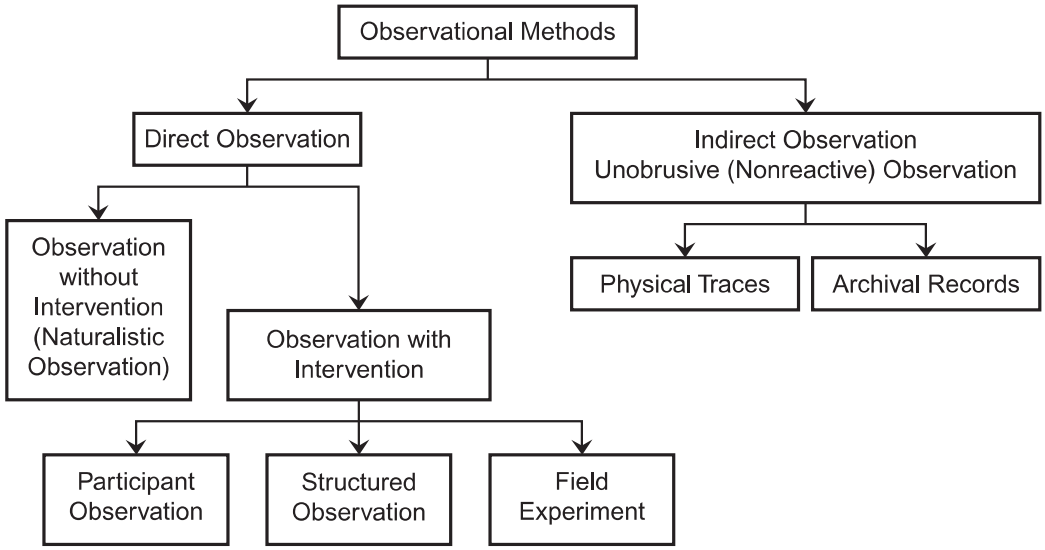
There are many situations where there may be more behavior going on than can be effectively observed. For example, if researchers observed students' food selections in the dining hall during peak hours, they would not be able to observe all the students. In this case, and in others like it, the researcher would use *subject sampling* to determine which students to observe. Similar to the procedures for time sampling, the researcher could either select students systematically (e.g., every 10th student) or select students randomly. In what is likely by now a familiar refrain, the goal of subject sampling is to obtain a representative sample, in this example, of all students eating in the dining hall.

Observational Methods

- Observational methods can be classified as direct observation or indirect observation.

Researchers often observe behavior while it occurs—that is, through *direct observation.* However, observations also can be made indirectly, as when researchers examine evidence of past behavior using physical traces or archival records. This is *indirect* (or unobtrusive) observation. Figure 4.1 illustrates the organization of observational methods. First we will discuss direct observational methods and then indirect (unobtrusive) methods.

FIGURE 4.1 Flow diagram of observational methods.



Direct Observational Methods

- Direct observational methods can be classified as "observation without intervention" or "observation with intervention."

When observing behavior directly, researchers make a decision regarding the extent to which they will intervene in the situation they observe. In this case, intervention refers to researchers' efforts to change or create the context for observation. The extent of intervention varies on a continuum from none (observation without intervention) to intervention that involves carrying out an experiment in a natural setting.

Observation without Intervention

- The goals of naturalistic observation are to describe behavior as it normally occurs and to examine relationships among variables.
- Naturalistic observation helps to establish the external validity of laboratory findings.
- When ethical and moral considerations prevent experimental control, naturalistic observation is an important research strategy.

Key Concept

Direct observation of behavior in a natural setting *without* any attempt by the observer to intervene is frequently called **naturalistic observation.** An observer using this method of observation acts as a passive recorder of events as they occur naturally. Although it is not easy to define a natural setting precisely (see Bickman, 1976), we can consider a natural setting one in which behavior ordinarily occurs and that has not been arranged specifically for the purpose of observing behavior. For example, Matsumoto and Willingham (2006) observed athletes in the "natural" (for these athletes) setting of an Olympic judo competition. Box 4.1 describes recent findings based on naturalistic observation within the field of ethology.

BOX 4.1

OBSERVATION: TAKING ANOTHER LOOK

Psychologists are not the only researchers who observe behavior in natural settings. Observation is a fundamental method in *ethology,* a branch of biology (Eibl-Eibesfeldt, 1975). Ethologists study the behavior of organisms in relation to their natural environment, typically logging countless hours of observation of animals in their natural settings. Speculations about the role of innate mechanisms in determining human behavior are not uncommon among ethologists.

For over a century, many biologists simply assumed that all animals engaged in female-male sex, without even looking at the sex of the animals. Recently, however, based on increasing numbers of observations from a large, diverse array of species, biologists suggest that same-sex sexual behavior is a nearly universal phenomenon (Bagemihl, 2000; Zuk, 2003). Biologists are taking another look at sex.

Researchers who study mating and procreation among animals have been struggling to interpret evidence indicating sexual and parenting behaviors among same-sex animals (Mooallem, 2010). Although most biologists avoid comparisons to human sexuality, the observations of same-sex behavior and co-parenting among animals has led to a great deal of controversy (see Figure 4.2). People on both sides of the sociopolitical debate regarding homosexuality have used evidence of same-sex behavior among animals to further their own agendas. A hallmark of scientific observation, however, is that it is objective and free from bias—including political agendas. Yet, many would wish to interpret animal sexuality using human terms, such as homosexuality or lesbianism, rather than to interpret the animal's behavior in its own context, with its own purpose.

The problem in understanding same-sex behaviors lies at the heart of evolutionary biology, namely, that all evolutionary-adaptive behavior is guided by a central goal: passing on genes. Nevertheless, biologists recently have developed theories suggesting that certain behaviors,

FIGURE 4.2 The children's book, *And Tango Makes Three* (Richardson & Parnell, 2005) is based on the story of two male penguins that were observed fostering a penguin chick at Central Park Zoo. The American Library Association reports that this was the most frequently banned book in 2009.

including sexual and parenting behaviors among same-sex animals, may be by-products of adaptation. This process of objective observation and theory construction forms the basis for all science. Yet, science, as we noted in Chapter 1, takes place in a cultural context that can lead some people to be less than objective when interpreting the results of this process.

Observing people in a psychological laboratory would not be considered naturalistic observation because a lab is created specifically to study behavior. Observation in natural settings often serves, among other functions, as a way to establish the external validity of laboratory findings—bringing the lab into the "real world." This is one goal of research conducted by researcher A.D.I. Kramer, who examines happiness using Facebook entries (*New York Times*, October 12, 2009). Observation of behavior in Internet discussion groups and chat rooms is yet another way that researchers have sought to describe behavior as it normally occurs (e.g., Whitlock, Powers, & Eckenrode, 2006). This recent form of "naturalistic" observation, however, raises the serious ethical issues that we discussed in Chapter 3 and will discuss later in this chapter (see also Kraut et al., 2004).

The major goals of observation in natural settings are to describe behavior as it ordinarily occurs and to investigate the relationship among variables that are present. Hartup (1974), for instance, chose naturalistic observation to investigate the frequency and types of aggression exhibited by preschoolers in a St. Paul, Minnesota, children's center. He distinguished hostile aggression (person-directed) from instrumental aggression (aimed at the retrieval of an object, territory, or privilege). Although he observed boys to be more aggressive overall than girls, his observations provided no evidence that the types of aggression differed between the sexes. Thus, Hartup was able to conclude that, with respect to hostile aggression, there was no evidence that boys and girls were "wired" differently.

Hartup's study of children's aggression illustrates why a researcher may choose to use naturalistic observation rather than to manipulate experimental conditions to study behavior. There are certain aspects of human behavior that moral or ethical considerations prevent us from controlling. For example, researchers are interested in the relationship between early childhood isolation and later emotional and psychological development. However, we would object strenuously if they tried to take children from their parents in order to raise them in isolation. Alternative methods of data collection must be considered if childhood isolation is to be investigated. For example, the effect of early isolation on later development has been studied through experimentation on animal subjects (Harlow & Harlow, 1966); observations of so-called feral children raised outside of human culture, presumably by animals (Candland, 1993); case studies of children subjected to unusual conditions of isolation by their parents (Curtiss, 1977); and systematic, direct observation of institutionalized children (Spitz, 1965). Moral and ethical sanctions also apply to investigating the nature of children's aggression. We would not want to see children intentionally harassed and picked on simply to record their reactions. However, as anyone who has observed children knows, there is plenty of naturally occurring aggression.

Hartup's study shows how naturalistic observation can be a useful method of gaining knowledge about children's aggression within moral and ethical constraints.

Observation with Intervention

- Most psychological research uses observation with intervention.
- The three methods of observation with intervention are participant observation, structured observation, and the field experiment.
- Whether "undisguised" or "disguised," participant observation allows researchers to observe behaviors and situations that are not usually open to scientific observation.
- If individuals change their behavior when they know they are being observed ("reactivity"), their behavior may no longer be representative of their normal behavior.
- Often used by clinical and developmental psychologists, structured observations are set up to record behaviors that may be difficult to observe using naturalistic observation.
- In a field experiment, researchers manipulate one or more independent variables in a natural setting to determine the effect on behavior.

It's not a secret. Scientists like to "tamper" with nature. They like to intervene in order to observe the effects and perhaps to test a theory. Intervention, rather than nonintervention, characterizes most psychological research. There are three important methods of observation that researchers use when they choose to intervene in natural settings: participant observation, structured observation, and the field experiment. The nature and degree of intervention varies across these three methods. We will consider each method in turn.

Key Concept

Participant Observation In **participant observation,** observers play a dual role: They observe people's behavior and they participate actively in the situation they are observing. In *undisguised* participant observation, individuals who are being observed know that the observer is present for the purpose of collecting information about their behavior. This method is used frequently by anthropologists who seek to understand the culture and behavior of groups by living and working with members of the group.

In *disguised* participant observation, those who are being observed do not know that they are being observed. As you might imagine, people do not always behave in the way they ordinarily would when they know their behavior is being recorded. As we'll discuss later in this chapter, a major problem when observing behavior is **reactivity.** Reactivity occurs when people react to the fact they are being observed by changing their normal behavior. Remember, researchers want to describe people's *usual* behavior. Therefore, researchers may decide to disguise their role as observers if they believe that people being observed will change their behavior once they know their activities are being recorded. Disguised participant observation raises ethical issues (e.g., privacy and informed consent) that must be addressed prior to implementing the study. We have considered these ethical issues in Chapter 3 and will discuss them further later in this chapter.

Key Concept

Participant observation allows an observer to gain access to a situation that is not usually open to scientific observation. For example, a researcher analyzing hate crimes against African Americans entered various "White racist Internet chat rooms" while posing as a "curious neophyte" (Glaser, Dixit, & Green, 2002). Such venues, of course, where violence is sometimes advocated, would normally not be open to scientific investigation.

In a classic study of psychiatric diagnosis and hospitalization of the mentally ill, Rosenhan (1973) employed disguised participant observers who sought admission to mental hospitals. Each complained of the same general symptom: That he or she was hearing voices. Most of the pseudopatients were diagnosed with schizophrenia. Immediately after being hospitalized, the participant observers stopped complaining of any symptoms and waited to see how long it took for a "sane" person to be released from the hospital. Once hospitalized, they recorded their observations. The researchers were hospitalized from 7 to 52 days, and when discharged, their schizophrenia was said to be "in remission." Apparently, once the pseudopatients were labeled schizophrenic, they were stuck with that label. There are, however, reasons to challenge this specific conclusion and other aspects of Rosenhan's (1973) study (see Box 4.2).

Because disguised participant observers have similar experiences as the people under study, they gain important insights and understanding of individuals or groups. The pseudopatients in the Rosenhan study, for instance, felt what it was like to be labeled schizophrenic and not to know how long it would be before they could return to society. An important contribution of Rosenhan's (1973) study was its illustration of the dehumanization that can occur in institutional settings.

A participant observer's role in a situation can pose serious problems in carrying out a successful study. Observers may, for instance, lose their scientific objectivity if they identify too closely with the people and situation they are observing. For example, a criminologist, Kirkham (1975), went through police academy training as an undisguised participant observer and became a uniformed patrol officer assigned to a high-crime area. His experiences as an officer led to unexpected and dramatic changes in his attitudes, personality, mood, and behavior. As Kirkham himself noted, he displayed "punitiveness, pervasive cynicism and mistrust of others, chronic irritability and free-floating hostility, racism, [and] a diffuse personal anxiety over the menace of crime and criminals" (p. 19). In situations such as these, participant observers must be aware of the threat to objective reporting due to their involvement in the situation, particularly as their involvement increases.

Another potential problem with participant observation is that the observer can influence the behavior of people being studied. It is likely that the participant observer will have to interact with people, make decisions, initiate activities, assume responsibilities, and otherwise act like everyone else in that situation. By participating in the situation, do observers change the participants and events? If people do not act as they normally would because of the participant observer, it is difficult to generalize results to other situations.

The extent of a participant observer's influence on the behavior under observation is not easily assessed. Several factors must be considered, such as whether participation is disguised or undisguised, the size of the group entered,

BOX 4.2

THINKING CRITICALLY ABOUT "ON BEING SANE IN INSANE PLACES"

In his article "On Being Sane in Insane Places," Rosenhan (1973) questioned the nature of psychiatric diagnosis and hospitalization. How could normal people be labeled as schizophrenic, one of the most severe mental illnesses we know? Why didn't the hospital staff recognize the pseudopatients were faking their symptoms? After days or weeks of hospitalization, why didn't the staff recognize that the pseudopatients were "sane," not insane?

These are important questions. After Rosenhan's research article was published in *Science* magazine, many psychologists and psychiatrists discussed and wrote articles in response to Rosenhan's questions (e.g., Spitzer, 1976; Weiner, 1975). Presented below are just a few of the criticisms of Rosenhan's research.

—We cannot criticize the staff for making a wrong diagnosis: A diagnosis based on faked symptoms will, of course, be wrong.
—The pseudopatients had more than one symptom; they were anxious (about being "caught"), reported they were distressed, and sought hospitalization. Is it "normal" to seek admission into a mental hospital?
— Did the pseudopatients really behave normally once in the hospital? Perhaps normal behavior would be to say something like, "Hey, I only pretended to be insane to see if I could be hospitalized, but really, I lied, and now I want to go home."
—Schizophrenics' behavior is not always psychotic; "true" schizophrenics often behave "normally." Thus, it's not surprising that the staff took many days to determine that the pseudopatients no longer experienced symptoms.
—A diagnosis of "in remission" was quite rare and reflects staff members' recognition that a pseudopatient was no longer experiencing symptoms. However, research on schizophrenia demonstrates that once a person shows signs of schizophrenia, he or she is more likely than others to experience these symptoms again. Therefore, the diagnosis of "in remission" guides mental health professionals as they try to understand a person's subsequent behavior.
—"Sane" and "insane" are legal terms, not psychiatric. The legal decision of whether someone is insane requires a judgment about whether a person knows right from wrong, which is irrelevant to this study.

As you can see, Rosenhan's research was controversial. Most professionals now believe that this study does not help us to understand psychiatric diagnosis. However, several important long-term benefits of Rosenhan's research have emerged:

—Mental health professionals are more likely to postpone a diagnosis until more information is gathered about a patient's symptoms; this is called "diagnosis deferred.''
—Mental health professionals are more aware of how their theoretical and personal biases may influence interpretations of patients' behaviors, and guard against biased judgments.
—Rosenhan's research illustrated the depersonalization and powerlessness experienced by many patients in mental health settings. His research influenced the mental health field to examine its practices and improve conditions for patients.

and the role of the observer in the group. When the group under observation is small or the activities of the participant observer are prominent, the observer is more likely to have a significant effect on people's behavior. This problem confronted several social psychologists who infiltrated a group of people who claimed to be in contact with beings from outer space (Festinger, Riecken, & Schachter, 1956). The group's leader said he had received a message from the aliens predicting a cataclysmic flood on a specific date. Because of the attitudes of members of the group toward "nonbelievers," the researchers were forced to make up bizarre stories in order to gain access to the group. This tactic worked

too well. One of the participant observers was even thought to be a spaceman bringing a message. The researchers had inadvertently reinforced the group's beliefs and influenced in an undetermined way the course of events that followed. As you are no doubt aware, a flood covering the entire northern hemisphere never occurred, but at least some of the group members came to use this disconfirmation as a means of strengthening their initial belief because their faith had prevented the prophesied flood. Thus, although participant observation may permit an observer to gain access to situations not usually open to scientific investigation, the observer using this technique must seek ways to deal with the possible loss of objectivity and the potential effects an observer may have on the behavior under study.

Structured Observation There are a variety of observational methods using intervention that are not easily categorized. These procedures differ from naturalistic observation because researchers intervene to exert some control over the events they are observing. The degree of intervention and control over events is less, however, than that seen in field experiments (which we describe briefly in the next section and in more detail in Chapter 6). We have labeled these procedures **structured observation.** Often the observer intervenes in order to cause an event to occur or to "set up" a situation so that events can be more easily recorded.

Key Concept

Researchers may create elaborate procedures to investigate a particular behavior fully. In a study of a phenomenon called inattentional blindness, researchers examined people's ability to notice unusual events while using a cell phone (Hyman, Boss, Wise, McKenzie, & Caggiano, 2009). Inattentional blindness occurs when people fail to notice new and distinctive stimuli in their environment, particularly when attention is focused elsewhere, such as a cell phone conversation. In their study the researchers used a *confederate*, that is, an individual in the research situation who is instructed to behave a certain way in order to create a situation for observing behavior. In Hyman et al.'s study, a confederate dressed as a clown rode a unicycle around a large sculpture in a central plaza area on a university campus (see Figure 4.3). Over a 1-hour period in which the clown was present, interviewers asked pedestrians who walked across the plaza whether they had seen anything unusual. If they answered yes, they were asked to specify what they had seen. If pedestrians did not mention the clown, they were asked specifically whether they had seen the unicycling clown.

This structured-observation procedure created the context for noting whether people are more likely to exhibit inattentional blindness while using a cell phone. The researchers classified pedestrians into one of four groups: cell phone user, single walker (with no electronics), walking singly while listening to music (e.g., using an MP3 player), or walking as a pair. Results indicated the cell phone users were least likely to notice the clown. Only 25% of cell phone users noticed the clown, compared to 51% of pedestrians walking alone, 61% of those listening to music, and 71% of individuals walking in pairs. Note that the individuals who might experience distractions due to music or walking with another person were more likely to notice the clown. This suggests that

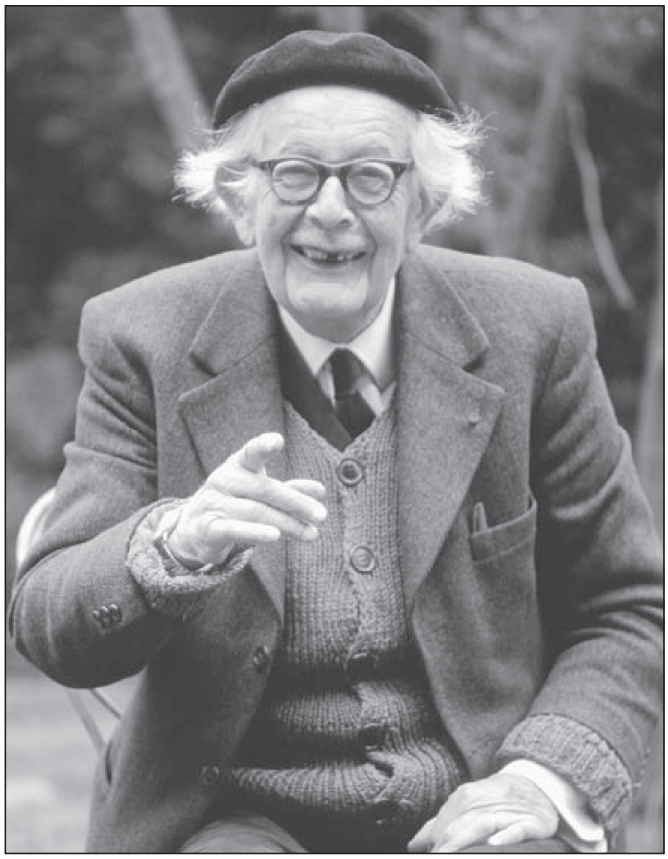
FIGURE 4.3 A photo of the unicycling clown in Hyman et al.'s (2009) study of inattentional blindness.

something particular about the divided attention when using a cell phone may be related to inattentional blindness. Hyman et al. (2009) note that if such a high degree of inattentional blindness is present during the simple activity of walking, the "blindness" that occurs with cell phone use may be much greater while driving a car.

Structured observations may be arranged in a natural setting, as in the Hyman et al. (2009) study, or in a laboratory setting. Clinical psychologists often use structured observations when making behavioral assessments of parent-child interactions. For example, researchers have observed play between mothers and children from maltreating (e.g., abusing, neglecting) families and nonmaltreating families (Valentino, Cicchetti, Toth, & Rogosch, 2006). Mothers were videotaped in a laboratory setting through a one-way mirror while interacting with their children in different contexts arranged by the researchers. In these structured observations, children from abusing families engaged in less independent play than children from nonmaltreating families and mothers in these families differed in their attention-directing behaviors. Valentino et al. suggest their study sheds light on the effect of a maltreating environment on children's social cognitive development, and they discuss implications for intervention.

Developmental psychologists frequently use structured observations. Jean Piaget (1896–1980) is perhaps most notable for his use of these methods (see Figure 4.4). In many of Piaget's studies, a child is first given a problem to solve

FIGURE 4.4 Jean Piaget (1896–1980) used structured observation to investigate children's cognitive development.



and then given several variations of the problem to test the limits of the child's understanding. These structured observations have provided a wealth of information regarding children's cognition and are the basis for Piaget's "stage theory" of intellectual development (Piaget, 1965).

Structured observation is a middle ground between the passive nonintervention of naturalistic observation and the systematic control and manipulation of independent variables in laboratory experiments. This compromise allows researchers to make observations in more natural settings than the laboratory. Nevertheless, there may be a price to pay. If observers fail to follow similar procedures each time they make an observation, it is difficult for other observers to obtain the same results when investigating the same problem. Uncontrolled, and perhaps unknown, variables may play an important part in producing the behavior under observation. To prevent this problem, researchers must be consistent in their procedures and try to "structure" their observations as similarly as possible across observations.

Key Concept

Field Experiments When a researcher manipulates one or more independent variables in a natural setting in order to determine the effect on behavior, the procedure is called a **field experiment.** The field experiment represents the most extreme form of intervention in observational methods. The essential difference

between field experiments and other observational methods is that researchers exert more control in field experiments when they manipulate an independent variable. Field experiments are frequently used in social psychology. For example, confederates have posed as robbers to investigate people's reaction to a crime, and researchers may manipulate the number of other bystanders (confederates) present to determine when people are most likely to help (Latané & Darley, 1970). Similarly, confederates have been used to cut into a waiting line in order to study those already in line (Milgram, Liberty, Toledo, & Wackenhut, 1986). In one field experiment, people's reactions to the intrusion were lessened when confederates also waited in line but did not object to the line cutting. Our discussion of experimental methods will continue in Chapter 6.

Indirect (Unobtrusive) Observational Methods

- An important advantage of indirect observational methods is that they are nonreactive.
- Indirect, or unobtrusive, observations can be obtained by examining physical traces and archival records.

STRETCHING EXERCISE

In this exercise we ask you to respond to the questions that follow this brief description of an observational study.

Students in a research methods class did an observational study to investigate whether students' ability to concentrate while studying was affected by where they studied. Specifically, students were observed in two locations on campus, the library and a lounge in the student union. The research methods students made their observations while appearing to be studying in the library or the lounge. They observed only students sitting alone in each location who had study materials such as a textbook or a notebook open in front of them. During a 5-minute observation period, the observers recorded the amount of time each student was studying, as indicated by either looking at the materials or writing. The student observers expected to find that students would be able to concentrate better in the library than in the student union.

Five student observers made observations for a total of 60 students in the library and 50 students in the student-union lounge from 9 to 11 P.M. on the same Monday evening. The mean time that students in the library spent studying was 4.4 of the 5.0 minutes. The corresponding mean time for students in the student union was 4.5 of the 5.0 minutes. The research methods students were surprised by two aspects of their findings. First, they were surprised to find that students studied for nearly 90% of the 5-minute study interval. They were even more surprised that, contrary to their prediction, the study times did not differ for the two locations.

1. Identify what type of observational method the students used in their study, and explain what characteristics of their study you used to make your identification.
2. How might the decision to use 5-minute observation periods affect the observers' ability to study concentration?
3. Why would the time-sampling plan in a study of this type be especially important? How could the time-sampling plan used in this study be improved to increase external validity?
4. Consider for the sake of this question that students can concentrate better in the library than in the student-union lounge. How could the nature of the material that the students were studying in the two locations have led to the finding that there was no difference between the observed concentration by students in the library and in the student union?

TABLE 4.1 INDIRECT (UNOBTRUSIVE) MEASURES

Physical Traces	Archival Records
1. ***Use traces:*** physical evidence that results from the use (or nonuse) of an item *Examples:* cans in a recycling bin, pages highlighted in a textbook, wear and tear on video game controllers	1. ***Running records:*** public and private documents that are produced continuously *Examples:* records for sports teams, Facebook and Twitter entries
2. ***Products:*** creations, constructions, or other artifacts of behavior *Examples:* petroglyphs (ancient rock paintings), MTV, *Harry Potter* action figures	2. ***Records for specific episodes:*** documents that describe specific events *Examples:* birth certificates, marriage licenses, college degrees

Based on distinctions made by Webb et al. (1981).

Key Concept

We have been discussing observational methods in which an observer directly observes and records behavior in a setting. However, behavior can also be observed indirectly through records and other evidence of people's behavior. These methods are often called **unobtrusive measures** because the researcher does not intervene in the situation and individuals are not aware of the observations. An important advantage of these methods is that they are *nonreactive*. A behavioral measure is reactive when participants' awareness of an observer affects the measurement process. Because unobtrusive observations are made indirectly, it is impossible for people to react, or change their behavior, while researchers observe. Unobtrusive methods also yield important information that can confirm or challenge conclusions based on direct observation, making these methods an important tool in the multimethod approach to research.

In this section we will describe these indirect methods, which involve the investigation of physical traces and archival records (see Table 4.1).

Physical Traces

- Two categories of physical traces are "use traces" and "products."
- Use traces reflect the physical evidence of use (or nonuse) of items and can be measured in terms of natural or controlled use.
- By examining products people own or the products produced by a culture, researchers test hypotheses about attitudes, preferences, and behavior.
- The validity of physical trace measures is examined by considering possible sources of bias and by seeking converging evidence.

Key Concept

As everyone who has read a few detective stories knows, examining physical evidence of past behavior can provide important clues about the characteristics of individuals and events. For example, the size of footprints in the ground says something about the size and age of the person who stepped there. The distance between footprints can indicate whether the person was walking or running. **Physical traces** are the remnants, fragments, and products of past behavior. Two categories of physical traces are "use traces" and "products."

Use traces are what the label implies—the physical evidence that results from the use (or nonuse) of an item. Remains of cigarettes in ashtrays, aluminum

cans in a recycling bin, and fingerprints on a murder weapon are examples of use traces. Clock settings are a use trace that may tell us about the degree to which people in different cultures are concerned with punctuality, and marks in textbooks may inform researchers which classes a students likes the best (or, at least, studies the most).

In addition, we can classify use traces according to whether the researcher intervenes while collecting data regarding the use of particular items. *Natural-use traces* are observed without any intervention by a researcher and reflect naturally occurring events. In contrast, *controlled-use traces* result from some intervention by a researcher. A study by Friedman and Wilson (1975) illustrates the distinction between these two types of use measures.

The investigators used both natural- and controlled-use traces to investigate college students' use of textbooks. Prior to the start of a course, they affixed tiny glue seals between pages of the textbooks. At the end of the course, they examined the textbooks to determine how many seals had been broken and where the broken seals were located. Because they controlled the presence of glue seals in the books, this would be an example of a controlled-use trace. These investigators also analyzed the frequency and nature of underlining in the textbooks, a natural-use measure because underlining is typically associated with textbook use. Analysis of both types of use traces indicated that students more often read the early chapters of the book than the later chapters.

Products are the creations, constructions, or other artifacts of behavior. Anthropologists often are interested in the surviving products of ancient cultures. By examining the types of vessels, paintings, tools, and other artifacts, anthropologists can describe patterns of behavior from thousands of years ago. Plenty of modern-day products provide insight into our culture and behavior, including television shows, music, fashion, and electronic devices. For instance, vehicle bumper stickers permit an acceptable outlet for the expression of public emotion and also allow individuals to reveal their identification with particular groups and beliefs (Endersby & Towle, 1996; Newhagen & Ancell, 1995). Tattoos and body piercings may function in a similar way in some cultures (see Figure 4.5).

The examination of products allows researchers to test important hypotheses about behavior. For example, psychologists examined food-related products in the United States and France to investigate the "French paradox" (Rozin, Kabnick, Pete, Fischler, & Shields, 2003). The term "French paradox" refers to the fact that obesity rates and the mortality rate from heart disease are much lower in France than the U.S., despite the fact that the French eat more fatty foods and fewer reduced-fat foods than Americans. Several hypotheses have been offered for these differences, ranging from metabolism differences, stress levels, and consumption of red wine. Rozin et al. hypothesized that the French simply eat less and they examined food products, specifically portion sizes, in both countries to test this hypothesis. They found that American restaurant portions were on average 25% greater than in comparable French restaurants, and that portion sizes on American supermarket shelves were generally larger. Their observation of products supported their hypothesis that the differences in obesity and mortality due to heart disease are because the French eat less than Americans.

FIGURE 4.5 Many cultures have used tattoos and body piercings as a means of self-expression and group identification.

The indirect observation of physical traces offers researchers valuable and sometimes innovative means to study behavior, and the measures available are limited only by the ingenuity and resourcefulness of the investigator. However, the validity of physical-trace measures must be examined carefully and verified through independent sources of evidence. Validity refers to the truthfulness of a measure and we must ask, as with all measurement, whether physical traces truthfully inform us about people's behavior.

Bias can be introduced in the way use traces are laid down and the manner in which traces survive over time. For example, does a well-worn path to the right in a museum indicate people's interest in objects in that direction or simply a natural human tendency to turn right? Does the number of cans found in recycling containers at a university reflect students' preferences for certain brands or simply what is available in campus vending machines? Does highlighting in a textbook reflect a particular student's study of the material or the accumulated use of the book over time by many students as the book is sold and resold? Do product sizes on supermarket shelves in America and France reflect different family sizes in the two countries or preferences for portion sizes? Whenever possible, researchers need to obtain supplementary evidence for the validity of physical traces (see Webb et al., 1981). Alternative hypotheses for observations of physical traces must be considered and care must also be taken when comparing results across studies to make sure that measures are defined similarly.

Archival Records

- Archival records are the public and private documents describing the activities of individuals, groups, institutions, and governments, and comprise running records and records of specific, episodic events.
- Archival data are used to test hypotheses as part of the multimethod approach, to establish the external validity of laboratory findings, and to assess the effects of natural treatments.
- Potential problems associated with archival records include selective deposit, selective survival, and the possibility of spurious relationships.

Consider for a moment all of the data about you that exist in various records: birth certificate; school enrollment and grades; credit/debit card purchases; driver's license, employment and tax records; medical records; voting history; e-mail, texting, and cell phone accounts; and if you're active on sites such as Facebook and Twitter, countless entries describing your daily life. Now multiply this by the millions of other people for whom similar records exist and you will only touch upon the amount of data "out there." Add to this all of the data available for countries, governments, institutions, businesses, media, and you will begin to appreciate the wealth of data available to psychologists to describe people's behavior using archival records.

Key Concept

Archival records are the public and private documents describing the activities of individuals, groups, institutions, and governments. Records that are continuously kept and updated are referred to as *running records*. The records of your academic life (e.g., grades, activities) are an example of running records, as are the continuous records of sports teams and the stock market. Other records, such as personal documents (e.g., birth certificates, marriage licenses), are more likely to describe specific events or episodes, and are referred to as *episodic records* (Webb et al., 1981).

As measures of behavior, archival data share some of the same advantages as physical traces. They are unobtrusive measures that are used to complement hypothesis testing based on other methods, such as direct observation, laboratory experiments, and surveys. When findings from these various approaches converge (or agree), the external validity of the findings increases. That is, we can say the findings *generalize* across the different research methods and enhance support for the hypothesis being tested. For example, recall the physical trace measures relating to portion size used to test the hypothesis concerning the "French paradox," namely, that the French eat less than Americans (Rozin et al., 2003). These researchers also examined archival records to test their hypothesis. They examined restaurant guides in two cities, Philadelphia and Paris, and recorded the number of references to "all-you-can-eat" buffets. Using an existing archival record (restaurant guides), they found converging evidence for their hypothesis: Philadelphia had 18 all-you-can-eat options and Paris had none.

Researchers may examine archives to assess the effect of a *natural treatment*. Natural treatments are naturally occurring events that significantly impact society or individuals. Because it is not always possible to anticipate these events, researchers who want to assess their impact must use a variety of behavioral

BOX 4.3

THE SCIENCE OF FREAKONOMICS

Do school teachers cheat on tests so that they and their students will look good?
Do police really lower crime rates?
Why does capital punishment not deter criminals?
Which is more dangerous to your child: the family owning a swimming pool or a gun?
Why are doctors so bad at washing their hands?
What's the best way to catch a terrorist?
Are people hard-wired for altruism or selfishness?
Why is chemotherapy prescribed so often if it's so ineffective?

These questions, and others, were asked by the maverick social scientist, Steven D. Levitt, in his best-selling books, *Freakonomics* and *SuperFreakonomics* (Levitt & Dubner, 2005; 2009). The answers he gives come from archival analyses of student test scores, sports records, crime statistics, birth and death statistics, and much more. We won't give away all the answers based on this clever researcher's mining of society's archives, but we will say that in this era of high-stakes testing, public school teachers sometimes cheat, and if you own both a gun and a swimming pool, your child is 100 times more likely to die by drowning than by gunplay.

measures, including archival data. Acts of terrorism such as 9/11, drastic economic events such as the worldwide economic collapse in 2008, and the enactment of new laws and reforms are examples of the kinds of events that may have important effects on behavior and can be investigated using archival data. Also, individuals experience naturally occurring events in their lives, such as death or divorce of parents, chronic illness, or relationship difficulties. The effects of these events can be explored using archival data. For example, a researcher may examine school records of absenteeism or grades to investigate children's responses to parental divorce. Similarly, Friedman et al. (1995) and Tucker et al. (1997) used archival data available from a longitudinal study begun in 1921 on a sample of 1,500 children. By also examining death records years later for individuals in the original sample, these investigators were able to determine that parental divorce was associated with earlier death for individuals in the study.

Researchers gain several practical advantages by using archival records. Archival data are plentiful and researchers can avoid an extensive data collection stage—data are simply waiting for researchers! Also, because archival information is often part of the public record and usually does not identify individuals, ethical concerns are less worrisome. As more and more archival sources become available through the Internet, researchers will find it even easier to examine behavior in this way (see Box 4.3).

Key Concept

Researchers, however, need to be aware of the problems and limitations of archival records. Two problems are *selective deposit* and *selective survival* (see Webb et al., 1981). These problems occur because there are biases in how archives are produced. **Selective deposit** occurs when some information is selected to be deposited in archives, but other information is not. For example, consider that great archive, the high school yearbook. Not all activities, events, and groups are selected to appear in the yearbook. Who decides what is prominently displayed

in the yearbook? When some events, activities, or groups have a better chance to be selected than others, bias exists. Or consider the fact that politicians and others who are constantly exposed to reporters know how to "use" the media by declaring that some statements are "off the record." This can be seen as a problem of selective deposit—only certain information is "for the record." You might also recognize this as a problem of reactivity, in that when deciding what is "for the record," individuals are reacting to the fact that their remarks are being recorded.

Interestingly, the *Congressional Record* is ostensibly a spontaneous record of speeches and remarks made before the Congress, but legislators actually have the opportunity to edit and alter the record before it is permanently recorded, and even to add documents into the record that were never read before Congress! No doubt remarks that are, in hindsight, less than politically expedient are changed prior to publication in the *Congressional Record.* This, too, is an example of selective deposit and can result in a biased account of the activities presented before Congress.

Key Concept

Selective survival arises when records are missing or incomplete (something an investigator may not even be aware of). Researchers must consider whether some records "survived," whereas others did not. Documents that are particularly damaging to certain individuals or groups may vanish, for example, during the change from one presidential administration to another. Family photo albums may "mysteriously" lose photos of individuals now divorced or photos from "fat years." In an archival study of letters printed in advice columns, Schoeneman and Rubanowitz (1985) cautioned that when analyzing the contents of the columns, they could not avoid the possibility of bias due to selective survival because advice columnists only print a fraction of the letters they receive; that is, only some of the letters "survived" to be printed in the newspaper column.

Another problem that can occur in the analysis of archival data is the possibility of identifying a *spurious relationship*. A spurious relationship exists when evidence falsely indicates that two or more variables are associated (see Chapter 5). False evidence can arise because of inadequate or improper statistical analyses, or more often, when variables are accidentally or coincidentally related. An association, or correlation, between two variables can occur when another, usually unidentified, third variable accounts for the relationship. For instance, archival records indicate that ice cream sales and crime rates are associated (as ice cream sales increase, so also do crime rates). Before we can conclude that eating ice cream causes people to commit crimes, it is important to consider that both variables, ice cream sales and crime rates, are likely affected by a third variable, seasonal temperatures. The spurious relationship between ice cream sales and crime rates can be accounted for by the third variable, temperature.

The possibility of biases due to selective deposit and selective survival, as well as spurious relationships, causes researchers to be appropriately cautious in reaching final conclusions based solely on the outcome of an archival study. Archival data are most useful when they provide complementary evidence in a multimethod approach to the investigation of a phenomenon.

Recording Behavior

- The goals of observational research determine whether researchers seek a comprehensive description of behavior or a description of only selected behaviors.
- How the results of a study are ultimately summarized, analyzed, and reported depends on how behavioral observations are initially recorded.

In addition to direct and indirect observation, observational methods also differ in the manner in which behavior is recorded. Sometimes researchers seek a *comprehensive* description of behavior and the situation in which it occurs. More often, they focus on only certain behaviors or events. Whether all behavior in a setting or only *selected* behaviors are observed depends on the researchers' goals. The important choice of how behavior is recorded ultimately determines how the results are measured, summarized, analyzed, and reported.

Comprehensive Records of Behavior

- Narrative records in the form of written descriptions of behavior, and audio and video recordings, are comprehensive records.
- Researchers classify and organize data from narrative records to test their hypotheses about behavior.
- Narrative records should be made during or soon after behavior is observed, and observers must be carefully trained to record behavior according to established criteria.

Key Concept

When researchers seek a comprehensive record of behavior, they often use narrative records. **Narrative records** provide a more or less faithful reproduction of behavior as it originally occurred. To create a narrative record, an observer can write descriptions of behavior, or use audio or video recordings. For example, videos were used to record the mother-child interactions among maltreating and nonmaltreating families (Valentino et al., 2006).

Once narrative records are created, researchers can study, classify, and organize the records to test their hypotheses or expectations about the behaviors under investigation. Narrative records differ from other forms of recording and measuring behavior because the classification of behaviors is done *after* the observations are made. Thus, researchers must make sure that the narrative records capture the information that will be needed to evaluate the hypotheses of the study.

Hartup (1974) obtained narrative records as part of his naturalistic study of children's aggression. Consider this sample narrative record from Hartup's study:

> Marian [a 7-year old] . . . is complaining to all that David [who is also present] had squirted her on the pants she has to wear tonight. She says, "I'm gonna do it to him to see how he likes it." She fills a can with water and David runs to the teacher and tells of her threat. The teacher takes the can from Marian. Marian attacks David and pulls his hair very hard. He cries and swings at Marian as the teacher tries to restrain him; then she takes him upstairs. . . . Later, Marian and Elaine go upstairs and into the room where David is seated with a teacher. He

> throws a book at Marian. The teacher asks Marian to leave. Marian kicks David, then leaves. David cries and screams, "Get out of here, they're just gonna tease me." (p. 339)

Hartup instructed his observers to use precise language when describing behavior and to avoid making inferences about the intentions, motives, or feelings of the participants. Note that we are not told why David might want to throw a book at Marian or how Marian feels about being attacked. Hartup believed that certain antecedent behaviors were related to specific types of aggression. By strictly excluding any references or impressions of the observers, individuals who examined the narrative would not be influenced by the observer's inferences. Thus, the content of the narrative records could be classified and coded in an objective manner.

Not all narrative records are as focused as those obtained by Hartup, nor do narrative records always avoid inferences and impressions of the observer. Narrative records also are not always meant to be comprehensive descriptions of behavior. For example, *field notes* include only the observer's running descriptions of the participants, events, settings, and behaviors that are of particular interest to the observer, and may not contain an exact record of everything that occurred. Field notes are used by journalists, social workers, anthropologists, psychologists, and others, and are probably used more frequently than any other kind of narrative record. Events and behaviors are likely to be interpreted in terms of the observer's specialized knowledge and field notes tend to be highly personalized (Brandt, 1972). For example, a clinical psychologist may record specific behaviors of an individual with knowledge of that individual's diagnosis or particular clinical issues. The usefulness of field notes as scientific records depends on the accuracy and precision of their content which, in turn, depend critically on the training of the observer and the extent to which the recorded observations can be verified by independent observers and through other means of investigation.

Practical and methodological considerations dictate the manner in which narrative records are made. *As a general rule, records should be made during or as soon as possible after behavior is observed.* The passage of time blurs details and makes it harder to reproduce the original sequence of actions. In addition, decisions regarding what should be included in a narrative record, the degree of observer inference, and the completeness of the narrative record must be decided prior to beginning a study (see, for example, Brandt, 1972). Once the content of narrative records is decided, observers must be trained to record behavior according to the criteria that have been set up. Practice observations may have to be conducted and records critiqued by more than one investigator before "real" data are collected.

Selected Records of Behavior

- When researchers seek to describe specific behaviors or events, they often obtain quantitative measures of behavior, such as the frequency or duration of its occurrence.
- Quantitative measures of behavior use one of four levels of measurement scales: nominal, ordinal, interval, and ratio.

- Rating scales, often used to measure psychological dimensions, are frequently treated as if they are interval scales even though they usually represent ordinal measurement.
- Electronic recording devices may be used in natural settings to record behavior, and pagers sometimes are used to signal participants to report their behavior (e.g., on a questionnaire).

Often researchers are interested only in certain behaviors or specific aspects of individuals and settings. They may have specific hypotheses about the behavior they expect and clear definitions of the behaviors they are investigating. In this type of observational study, researchers typically measure the occurrence of the specific behavior while making their observations. For example, in their study of inattentional blindness, Hyman and his colleagues (2009) selected the behavior of whether people noticed the clown and quantified the number of people who noticed or did not notice the clown.

Suppose you wish to observe people's reactions to individuals with obvious physical disabilities using naturalistic observation. First you would need to define who is a "physically disabled person" and what constitutes a "reaction" to such a person. Are you interested in helping behaviors, approach/avoidance behaviors, eye contact, length of conversation, or in another physical reaction? Next you would need to decide how to measure these behaviors. Assume you choose to measure people's reactions by observing eye contact between individuals with and without physical disabilities. Exactly how should you measure eye contact? Should you simply measure whether an individual does or does not make eye contact, or do you want to measure the duration of any eye contact? Your decisions will depend on the hypotheses or goals of your study, and will be influenced by information gained by reading previous studies that used the same or similar behavioral measures. Unfortunately, previous research indicates that reactions to physically disabled individuals frequently can be classified as unfavorable (Thompson, 1982).

Key Concept

Measurement Scales When researchers decide to measure and quantify specific behaviors they must decide what scale of measurement to use. There are four levels of measurement, or **measurement scales,** that apply to both physical and psychological measurement: nominal, ordinal, interval, and ratio. The characteristics of each measurement scale are described in Table 4.2, and a detailed

TABLE 4.2 CHARACTERISTICS OF MEASUREMENT SCALES

Type of Scale	Operations	Objective
Nominal	Equal/not equal	Sort stimuli into discrete categories
Ordinal	Greater than/less than	Rank-order stimuli on a single dimension
Interval	Addition/multiplication/ subtraction/division	Specify the distance between stimuli on a given dimension
Ratio	Addition/multiplication subtraction/division/ formation of ratios of values	Specify the distance between stimuli on a given dimension and express ratios of scale values

BOX 4.4

MEASUREMENT "ON THE LEVEL"

The lowest level of measurement is called a *nominal scale;* it involves categorizing an event into one of a number of discrete categories. For instance, we could measure the color of people's eyes by classifying them as "brown-eyed" or "blue-eyed." When studying people's reactions to individuals with obvious physical disabilities, a researcher might use a nominal scale by measuring whether participants make eye contact or do not make eye contact with someone who has an obvious physical disability.

Summarizing and analyzing data measured on a nominal scale is limited. The only arithmetic operations that we can perform on nominal data involve the relationships "equal" and "not equal." A common way of summarizing nominal data is to report frequency in the form of proportion or percent of instances in each of the several categories.

The second level of measurement is called an ordinal scale. An *ordinal scale* involves ordering or ranking events to be measured. Ordinal scales add the arithmetic relationships "greater than" and "less than" to the measurement process. The outcome of a race is a familiar ordinal scale. When we know that an Olympic distance runner won a silver medal, we know the runner placed second but we do not know whether she finished second in a photo finish or trailed 200 meters behind the gold medal winner.

The third level of measurement is called an interval scale. An *interval scale* involves specifying how far apart two events are on a given dimension. On an ordinal scale, the difference between an event ranked first and an event ranked third does not necessarily equal the distance between those events ranked third and fifth. For example, the difference between the finishing times of the first- and third-place runners may not be the same as the difference in times between the third- and fifth-place runners. On an interval scale, however, differences of the same numerical size in scale values are equal. For example, the difference between 50 and 70 correct answers on an aptitude test is equal to the difference between 70 and 90 correct answers. What is missing from an interval scale is a meaningful zero point. For instance, if someone's score is zero on a verbal aptitude test, he or she would not necessarily have absolutely zero verbal ability (after all, the person presumably had enough verbal ability to take the test). Importantly, the standard arithmetic operations of addition, multiplication, subtraction, and division can be performed on data measured on an interval scale. Whenever possible, therefore, psychologists try to measure psychological dimensions using at least interval scales.

The fourth level of measurement is called a ratio scale. A *ratio scale* has all the properties of an interval scale, but a ratio scale also has an absolute zero point. In terms of arithmetic operations, a zero point makes the ratio of scale values meaningful. For example, temperature as expressed on the Celsius scale represents an interval scale of measurement. A reading of 0 degrees Celsius does not really mean absolutely no temperature. Therefore, it is not meaningful to say that 100 degrees Celsius is twice as hot as 50 degrees, or that 20 degrees is three times colder than 60 degrees. On the other hand, the Kelvin scale of temperature does have an absolute zero, and the ratio of scale values can be meaningfully calculated. Physical scales measuring time, weight, and distance can usually be treated as ratio scales. For example, someone who is 200 pounds weighs twice as much as someone who weighs 100 pounds.

description of measurement scales is provided in Box 4.4. You will need to keep these four measurement scales in mind as you select statistical procedures for analyzing the results of a research study. How data are analyzed depends on the measurement scale used. In this section we describe how the measurement scales can be used in observational research.

A *checklist* is often used to record nominal scale measures. To return to our example, an observer could record on a checklist whether individuals make eye contact or do not make eye contact with a physically disabled person, representing two discrete categories of behavior (a nominal measure). Checklists often include space to record observations regarding characteristics of participants, such as their race, sex, and age, as well as characteristics of the setting, such as time of day, location, and whether other people are present. Researchers typically are interested in observing behavior as a function of these participant and context variables. For example, Hyman et al. (2009) classified pedestrians in their study of inattentional blindness into four categories based on whether they were walking alone or in pairs and whether they were using a cell phone or music player (note that other categories, such as people walking in groups of three or more, were excluded).

The second level of measurement, an ordinal scale, involves ordering or ranking observations. Tassinary and Hansen (1998) used ordinal measurement to test a specific prediction of evolutionary psychology, namely, that female attractiveness is based on physical cues that simultaneously signal attractiveness *and* reproductive potential. The specific measure in this theory is the waist-to-hip ratio, with hips wider than waist indicating greater reproductive potential. In their study, undergraduates rank-ordered line drawings of female figures that varied in terms of height, weight, and hip size. That is, they ordered the drawings from least attractive to most attractive. Contrary to the prediction based on evolutionary psychology, physical attractiveness of the figures was directly and negatively related only to hip size, not the waist-to-hip ratio. Drawings with wider hips were more likely to be rank-ordered as lower in attractiveness.

In order to quantify behavior in an observational study, observers sometimes make *ratings* of behaviors and events based on their subjective judgments about the degree or quantity of some trait or condition (see Brandt, 1972). For example, Dickie (1987) asked trained observers to rate parent-infant interactions in a study designed to assess the effect of a parent training program. Observers visited the home and asked parents to "act as normal as possible—just as if we [the observers] weren't here." Observers made ratings using 7-point scales on 13 dimensions describing characteristics of verbal, physical, and emotional interaction. Ratings of 1 represented the absence or very little of the characteristic, and larger numbers represented increasing amounts of the trait. An example of one dimension, "parent's warmth and affection toward infant" is described in Table 4.3. Note that precise verbal descriptions are given for the four odd-numbered scale values to help the observers define different degrees of this trait. The even-numbered values (2, 4, 6) are used by observers to rate behaviors that they judge fall between the defined values. Based on observers' ratings, parents who took part in the program aimed at helping them to deal with their infant were rated higher on many of the 13 parent-child interaction variables than were parents who did not participate in the program.

At first glance, a rating scale such as the one in Table 4.3 appears to represent an interval scale of measurement—there is no true zero and the intervals between numbers appear to be equal. Closer examination, however, reveals that

TABLE 4.3 EXAMPLE OF RATING SCALE USED TO MEASURE A PARENT'S WARMTH AND AFFECTION TOWARD AN INFANT CHILD*

Scale Value	Description
1	There is an absence of warmth, affection, and pleasure. Excessive hostility, coldness, distance, and isolation from the child are predominant. Relationship is on an attacking level.
2	
3	There is occasional warmth and pleasure in interaction. Parent shows little evidence of pride in the child, or pride is shown in relation to deviant or bizarre behavior by the child. Parent's manner of relating is contrived, intellectual, not genuine.
4	
5	There is moderate pleasure and warmth in the interaction. Parent shows pleasure in some areas but not in others.
6	
7	Warmth and pleasure are characteristic of the interaction with the child. There is evidence of pleasure and pride in the child. Pleasure response is appropriate to the child's behavior.

*From materials provided by Jane Dickie.

most rating scales used by observers to evaluate people or events on a psychological dimension really yield only *ordinal* information. For a rating scale to be truly an interval level of measurement, a rating of 2, for instance, would have to be the same distance from a rating of 3 as a rating of 4 is from 5 or a rating of 6 is from 7. It is highly unlikely that human observers can make subjective judgments of traits such as warmth, pleasure, aggressiveness, or anxiety in a manner that yields precise interval distances between ratings. *However, most researchers assume an interval level of measurement when they use rating scales.* Deciding what measurement scale applies for any given measure of behavior is not always easy. If you are in doubt, you should seek advice from knowledgeable experts so that you can make appropriate decisions about the statistical description and analysis of your data.

Checklists also can be used to measure the *frequency* of particular behaviors in an individual or group over a period of time. The presence or absence of specific behaviors is noted at the time of each observation. After all the observations are made, researchers add up the number of times a particular behavior occurred. In these situations, frequency of responding can be assumed to represent a ratio level of measurement. That is, if "units" of some behavior (e.g., occasions when a child leaves a classroom seat) are counted, then zero represents the absence of that specific behavior. Ratios of scale values also would be meaningful. For example, a child who leaves her seat 20 times would have exhibited the target behavior twice as much as a child who leaves his seat 10 times.

Electronic Recording and Tracking Behavior also can be measured using electronic recording and tracking devices. For example, as part of a study investigating the relationship between cognitive coping strategies and blood pressure among college students, participants wore an ambulatory blood pressure

monitor on two "typical" school days, including a day with an exam (Dolan, Sherwood, & Light, 1992). Participants also completed questionnaires about their coping strategies and daily activities. The researchers compared blood pressure readings for different times of the day and as a function of coping style. Students who exhibited "high self-focused coping" (e.g., "keep to themselves and/or blame themselves in stressful situations," p. 233) had higher blood pressure during and after an exam than did those who did not use self-focused coping strategies.

Another electronic method is the "Internet daily diary" in which participants log on daily to a secure Internet site (with e-mail reminders) to report on daily events. Park, Armeli, and Tennen (2004) used this method to examine college students' moods and coping. Each day, students reported their most stressful event and how they coped with it. Results of this study indicated that positive moods were linked more with problem-focused coping strategies than with avoidance strategies, especially when the stressful events were perceived as controllable. Other researchers have asked participants to carry hand-held computers and to make "electronic diary" notes when prompted (e.g., McCarthy, Piasecki, Fiore, & Baker, 2006; Shiffman & Paty, 2006). Undoubtedly, as Internet access with cell phones becomes commonplace, electronic methods for data collection increasingly will be used by researchers.

Electronic recording methods often rely on participants' self-reports of mood and activities, not on direct observation of their behavior. As such, it is important that researchers devise techniques to detect biases in data collection (e.g., possible misrepresentation or omission of activities; see Larson, 1989, for a discussion of possible biases). These problems can be weighed against the time and costs sometimes required to obtain a comprehensive description of behavior using direct observation (e.g., Barker, Wright, Schoggen, & Barker, 1978).

Analysis of Observational Data

- Researchers choose qualitative data analysis or quantitative data analysis to summarize observational data.

After recording their observations of behavior, researchers analyze observational data in order to summarize people's behavior and to determine the reliability of their observations. The type of data analysis that researchers choose depends on the data they've collected and the goals of their study. For example, when researchers record selected behaviors using a measurement scale, the preferred data analysis is quantitative (i.e., statistical summaries and analyses). When comprehensive narrative records are obtained, researchers may choose either quantitative or qualitative analyses. We will describe qualitative analyses first.

Qualitative Data Analysis

- Data reduction is an important step in the analysis of narrative records.
- Researchers code behaviors according to specific criteria, for example, by categorizing behaviors.

- Content analysis is used to examine archival records and includes three steps: identifying a relevant source, sampling sections from the source, and coding units of analysis.

Analysis of Narrative Records Observational studies that use comprehensive narrative records or archival records provide a wealth of information–sometimes piles and piles of papers, video and audio recordings. Once the data are collected, how do researchers summarize all of this information? An important step in analyzing the content of narrative records is **data reduction,** the process of abstracting and summarizing behavioral data. In *qualitative* data analysis, researchers seek to provide a verbal summary of their observations and to develop a theory that explains behavior in the narrative records (see Miles & Huberman, 1994; Strauss & Corbin, 1990). In qualitative analysis, data reduction occurs when researchers verbally summarize information, identify themes, categorize and group pieces of information, and record their own observations about the narrative records.

Key Concept

Data reduction often involves the process of **coding,** which is the identification of units of behavior or particular events according to specific criteria that are related to the goals of the study. For example, in a study of preschool children, McGrew (1972) developed coding schemes to classify 115 different patterns of behavior according to the body part involved, ranging from facial expressions such as bared teeth, grin face, and pucker face, to locomotion behaviors such as gallop, crawl, run, skip, and step. Observers used the coding schemes to classify these behavioral patterns while they watched videos of children in preschool. Data reduction in this way (i.e., from videos to coded behaviors) allows researchers to determine relationships between specific types of behavior and the events that are antecedents of these behaviors. For example, McGrew found that children exhibit a "pout face" after losing a fight over a toy. McGrew also studied young chimpanzees and noted that these animals show a pout face when seeking reunion with their mother. Just after being frustrated (and often just prior to weeping), children exhibited a "pucker face." Interestingly, there was no record of a pucker face in the nonhuman primates.

Key Concept

Content Analysis of Archival Records As with narrative records, the amount of data obtained from archival records can be daunting, and the researcher's first step involves data reduction. In the simplest cases, only data reduction may be necessary. For example, a simple tally of votes by legislators on a particular issue may quickly and effectively summarize data in a government record. In many cases, however, gleaning relevant data from an archival source can require careful procedures and relatively complex analysis of the source's content.

Key Concept

Content analysis can be generally defined as any objective coding technique that allows researchers to make inferences based on specific characteristics in archival records (Holsti, 1969). Although content analysis is associated primarily with written communications, it may be used with any form of communication, including television and radio programs, speeches, films, interviews, and Internet content (including text and e-mail messages, "tweets," etc.). When television or radio broadcasts are studied, time is often used as a unit of *quantitative*

measurement (e.g., the amount of time members of different ethnic groups appear on screen). When the communication is written, quantitative analysis may examine single words, characters, sentences, paragraphs, themes, or particular items (Holsti, 1969). For example, researchers studying the quality of a marital relationship may count the couple's use of pronouns (*we, you, I, he,* and *she*) found in transcripts of their interactions (e.g., Simmons, Gordon, & Chambless, 2005). When newspaper content is analyzed, a frequently used quantitative measure is space—for instance, the number of column inches devoted to particular topics. Qualitative data analysis of archival records using content analysis is similar to the methods described for narrative records.

The three basic steps of content analysis for archival records include identifying a relevant source, sampling selections from the source, and coding units of analysis. A relevant archival source is one that allows researchers to answer the research questions of the study. Although researchers can be quite ingenious when identifying their source, often the identification of the archival source is relatively straightforward, as, for example, when researchers investigated the relationship between the likelihood of being sentenced to death and the extent to which defendants had a stereotypical Black appearance (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). They used as their archival source an extensive database of death-eligible cases from the state of Pennsylvania that contained prisoners' photographs, crime data, and sentencing outcomes. Their results indicated a disturbing outcome: Defendants who appeared more stereotypically Black (based on independent ratings) were more likely to receive the death sentence than those with less stereotypical features.

The second step in content analysis involves sampling appropriately from the archival source. Many databases and archival sources are so extensive that it would be impossible for an investigator to analyze all of the information in the source; therefore, the investigator must select some of the data with the goal of obtaining a representative sample. Ideally, a researcher would use some technique for randomly selecting portions of the archive. The extent to which the results of an archival study can be generalized (external validity) depends on the representativeness of the sample. Earlier we mentioned the results of an archival study that examined the relationship between parental divorce and premature mortality (Friedman et al., 1995). The sample of data for this archival study was based on a sample of children initially studied in 1921; clearly, it was not a random sample of divorce and mortality statistics. We might question the external validity of findings for the impact of parental divorce on children's lives at the beginning of the 20th century when divorce was less frequent and less socially acceptable. Very different findings may be observed today.

The last step in performing a content analysis is *coding*. This step requires that relevant descriptive categories and appropriate units of measure be defined (see Holsti, 1969). As with the choice of the archival source itself, the descriptive categories depend on the goals of the study. In order for coders to make reliable judgments about the archival data, they must be carefully trained and precise operational definitions must be used. For example, in a study of adolescents' self-injury behaviors, researchers used a set of binary (present/absent) codes to analyze the content of Internet message boards related to adolescent self-injury

(Whitlock, Powers, & Eckenrode, 2006). They derived their codes from interviews with self-injurers and from observations of messages posted on the Internet. They then examined 3,219 Internet postings from 10 Internet message boards over a 2-month period and coded, or categorized, the content into different themes, such as motivation for self-injury and methods of concealing their behavior. Similar to analysis of narrative records, data reduction using coding allows researchers to determine relationships between specific types of behavior and the events that are antecedents of these behaviors. Whitlock and her colleagues, for instance, identified "triggers" of self-injury behaviors in their coding and were able to identify the proportion of messages that described each trigger. Based on their coding, they observed that "conflict with important others" was the most frequent trigger (34.8%) of self-injury. By counting the occurrence of these triggers, these investigators moved from qualitative coding of the data to quantitative data analysis.

Quantitative Data Analysis

- Data are summarized using descriptive statistics such as frequency counts, means, and standard deviations.
- Interobserver reliability refers to the extent to which independent observers agree in their observations.
- Interobserver reliability is increased by providing clear definitions about behaviors and events to be recorded, by training observers, and by providing feedback about the accuracy of observations.
- High interobserver reliability increases researchers' confidence that observations about behavior are accurate (valid).
- Interobserver reliability is assessed by calculating percentage agreement or correlations, depending on how the behaviors were measured and recorded.

The goal of *quantitative data analysis* is to provide a numerical, or quantitative, summary of observations in a study. An important step is to calculate descriptive statistics that summarize the observational data, such as relative frequency, means, and standard deviations. Another important aspect of analyzing observational data is assessing the reliability of the observations. Unless the observations are reliable, they are unlikely to tell us anything meaningful about behavior. We will describe each of these aspects of quantitative data analysis in turn.

Descriptive Statistics The type of descriptive statistics used to summarize observational data depends on the scale of measurement used to record the data. As we saw, a nominal scale of measurement is used when behaviors and events are classified into mutually exclusive categories. Because a frequently used nominal measurement is whether a behavior is present or absent, the most common descriptive statistic for the nominal scale is *relative frequency*. To calculate a relative frequency the number of times a behavior or event occurs is tallied and then divided by the total number of observations. Relative frequency measures are expressed as either a proportion or a percentage (by multiplying the proportion by 100). We mentioned earlier that Whitlock and her colleagues

coded triggers for self-injury behavior among adolescents, with the most frequent trigger being "conflict with important others." They counted 212 mentions of conflict among the 609 messages in which triggers were mentioned. The relative frequency, then, is .348 (212 ÷ 609), or 34.8% of the messages.

When describing ordinal data, researchers often report the item most frequently ranked first among a set of items. For example, in surveys addressing citizens' concerns about the country, researchers may ask people to rank order items such as the economy, wars, education, environment, national security, and so forth, in terms of the priority for government action. When reporting the results, researchers may describe an item according to the percentage of people who ranked it first, such as "35% of respondents ranked the economy as their top priority for government action" (hypothetical data). A more complete description would include the percentage of first-rankings for the remaining items, such as "28% of respondents indicated the environment is their top priority, 25% indicated that wars are their top priority," and so on. Another way to describe ordinal data focuses on describing the percentages of 1st, 2nd, and 3rd, etc. rankings for a particular item selected from among the set of items. Hypothetically, this might appear as "35% of respondents ranked the economy as 1st priority, 25% of respondents ranked the economy as their 2nd in priority, 12% ranked it 3rd," and so on.

Different—and more informative—descriptive statistics are reported when behavior is recorded on at least an interval scale of measurement. One or more measures of central tendency are used when observations are recorded using interval-scale ratings or when ratio-scale measures of time (duration, latency) are used. The most common measure of central tendency is the *arithmetic mean*, or *average*. The mean describes the "typical" score in a group of scores and provides a useful measure to summarize the performance of an individual or group. For a more complete description of performance, researchers also report a measure of variability or dispersion of scores around the mean. The *standard deviation* approximates the average distance of a score from the mean.

Now may be a good time to review measures of central tendency and variability, as well as general guidelines for systematically analyzing data sets. The first few pages of Chapter 11 are devoted to these issues.

LaFrance and Mayo (1976) reported means and standard deviations in their study of eye contact between same-race pairs of Black and White people in conversation. The number of seconds that each listener in a pair spent looking into the speaker's face was recorded. Table 4.4 gives the means and standard deviations summarizing the results of this study. The means indicate that on average, White listeners spent more time looking into the faces of speakers than did Black listeners. This finding was obtained for both same-sex pairs and male-female pairs. The standard deviations indicate that male pairs showed less variability than either female pairs or male-female pairs. Measures of central tendency and variability provide a remarkably efficient and effective summary of the large numbers of observations that were made in this study.

TABLE 4.4 MEANS AND STANDARD DEVIATIONS DESCRIBING THE TIME (IN SECONDS) THAT LISTENERS SPENT LOOKING INTO THE FACE OF A SAME-RACE SPEAKER PER 1-MINUTE OBSERVATION UNIT*

Group	Mean	Standard Deviation
Black conversants		
Male pairs	19.3	6.9
Female pairs	28.4	10.2
Male–female pairs	24.9	11.6
White conversants		
Male pairs	35.8	8.6
Female pairs	39.9	10.7
Male–female pairs	29.9	11.2

*From LaFrance and Mayo (1976).

Key Concept

Observer Reliability In addition to descriptive statistics, researchers examine the extent to which the observations in their study are reliable. You may recall that reliability refers to consistency, and an analysis of reliability in an observational study asks if independent observers viewing the same events would obtain the same results. The degree to which two (or more) independent observers agree is referred to as **interobserver reliability.** When observers disagree, we become uncertain about what is being measured and the behaviors and events that actually occurred. Low interobserver reliability is likely to result when the event to be recorded is not clearly defined and observers are left to their own subjective judgments to make decisions about behavior. In addition to providing precise verbal definitions to improve reliability among observers, researchers can give concrete examples, including photographs and videos of specific behaviors to be observed. Interobserver reliability is also generally increased by training observers and giving them opportunities to practice making their observations. It is especially helpful during the training and practice to give observers specific feedback regarding any discrepancies between their observations and those of others (Judd, Smith, & Kidder, 1991).

Highly reliable observations do not necessarily mean the observations will be accurate. Consider two observers who reliably agree about what they saw but both are "in error" to the same degree. Neither observer would be providing an accurate record of behavior. For example, both might be influenced in a similar way by what they expect the results of their observation to be. Instances are occasionally reported in the media of several observers claiming to see the same thing (such as an unidentified flying object, or UFO), only to have the event turn out to be something else (a weather balloon). Nevertheless, when two independent observers agree, we are generally more inclined to believe that their observations are accurate and valid than when data are based on the observations of a single observer. In order for observers to be independent, each must be unaware of what the other has recorded. The chance of both observers being influenced to the same degree by expectancies, fatigue, or boredom is generally small enough that we can be confident that what they agree upon in their reports actually occurred. Of course, the more independent observers agree, the more confident we become.

The way in which interobserver reliability is assessed depends on how behavior is measured. When events are classified according to mutually exclusive categories (nominal scale), observer reliability is generally assessed using a percentage agreement measure. A formula for calculating percentage agreement between observers is

$$\frac{\text{Number of times two observers agree}}{\text{Number of opportunities to agree}} \times 100$$

In his study of childhood aggression, Hartup (1974) reported measures of reliability using percentage agreement that ranged from 83% to 94% for observers who coded type of aggression and the nature of antecedent events in narrative records. Although there is no hard-and-fast rule that defines low interobserver reliability, researchers generally report estimates of reliability that exceed 85% in the published literature, suggesting that percentage agreement much lower than that is unacceptable.

In many observational studies, data are collected by several observers who observe at different times. Under these circumstances, researchers select a sample of the observations to measure reliability. For example, two observers might record behavior according to time-sampling procedures and observe at the same time for only a subset of times. The percentage agreement for the times in which both observers are present can be used to estimate the degree of reliability for the study as a whole.

When data are measured using an ordinal scale, the Spearman rank-order correlation is used to assess interobserver reliability. When observational data are measured on an interval or ratio scale, such as when time is the measured variable, observer reliability can be assessed using a Pearson Product-Moment Correlation Coefficient, *r*. For example, LaFrance and Mayo (1976) obtained measures of reliability when observers recorded how much time a listener gazed into the speaker's face during a conversation. Observer reliability in their study was good; they found an average correlation of .92 between pairs of observers who recorded time spent in eye contact.

Key Concept

A *correlation* exists when two different measures of the same people, events, or things vary together—that is, when scores on one variable covary with scores on another variable. A **correlation coefficient** is a quantitative index of the degree of this covariation. When observation data are measured using interval or ratio scales, a Pearson correlation coefficient, *r*, may be used to obtain a measure of interobserver reliability. The correlation tells us how well the ratings of two observers agree.

The correlation coefficient indicates the *direction* and *strength* of the relationship. Direction can be either positive or negative. A positive correlation indicates that as the values for one measure increase, the values of the other measure also increase. For example, measures of smoking and lung cancer are positively correlated. A negative correlation indicates that as the values of one measure increase, the values of the second measure decrease. For instance,

time spent watching television and scores on academic tests are negatively correlated. When assessing interobserver reliability, researchers seek positive correlations.

The strength of a correlation refers to the degree of covariation present. Correlations range in size from −1.00 (a perfect negative relationship) to 1.00 (a perfect positive relationship). A value of 0.0 indicates there is no relationship between the two variables. The closer a correlation coefficient is to 1.0 or −1.0, the stronger the relationship between the two variables. Note that the sign of a correlation signifies only its direction; a correlation coefficient of −.46 indicates a stronger relationship than one that is .20. We suggest that measures of interobserver reliability that exceed .85 indicate good agreement between observers (but the higher, the better!).

In Chapter 5 we discuss the use of correlations for making predictions. In addition, Chapter 11 provides a detailed discussion of correlations, including how relationships between two variables can be described graphically using scatterplots, how Pearson Product-Moment Correlation Coefficients are computed, and how these correlations are best interpreted. If you want to become more familiar with the topic of correlation, refer to Chapter 11.

Thinking Critically About Observational Research

A good observational study involves choosing how to sample behavior and events to observe, selecting the appropriate observational method, and deciding how to record and analyze observational data. Now that you know the basics of observational methods, you also need to know about potential problems that can occur. The first problem is associated with the influence of the observer on behavior; a second problem occurs when observers' biases influence what behavior they choose to record. We'll consider each of these problems in turn.

Influence of the Observer

- The problem of reactivity occurs when the observer influences the behavior being observed.
- Research participants may respond to demand characteristics in the research situation to guide their behavior.
- Methods to control reactivity include concealing the observer's presence, adaptation (habituation, desensitization), and indirect observation (physical traces, archival records).
- Researchers must consider ethical issues when attempting to control reactivity.

Reactivity The presence of an observer can lead people to change their behavior because they know they are being observed. We first addressed this issue of *reactivity* in the section describing participant observation. When individuals "react" to the presence of an observer, their behavior may not represent

their typical behavior—that is, their behavior when an observer is not present. Underwood and Shaughnessy (1975) relate how a student, as part of a class assignment, set out to observe whether drivers came to a complete stop at an intersection with a stop sign. The observer positioned himself on the street corner with clipboard in hand, and soon noticed that all of the drivers stopped at the stop sign. His presence influenced their behavior. When he concealed himself near the intersection, he found that drivers' behavior changed and he was able to gather data for his study.

Research participants can respond in very subtle ways when they are aware that their behavior is being observed. For instance, participants are sometimes apprehensive and anxious about participating in psychological research, and measures of arousal (e.g., heart rate) may change simply because an observer is present. Research participants who wear an electronic beeper that signals them to record their behavior and mood also can be expected to change their behavior (e.g., Larson, 1989).

Individuals often react to the presence of an observer by trying to behave in ways they think the researcher wants them to behave. Knowing they are part of a scientific investigation, individuals usually want to cooperate and be "good" participants. Research participants often try to guess what behaviors are expected, and they may use cues and other information to guide their behavior (Orne, 1962). These cues in the research situation are called **demand characteristics.** Orne suggests that individuals generally ask themselves the question, "What am I supposed to be doing here?" To answer this question, participants pay attention to the cues present in the setting, the research procedure, and implicit cues given by the researcher. To the extent that participants change their behavior as they pay attention to demand characteristics, the external validity of the research is threatened. The ability to generalize the research findings (external validity) is threatened when research participants behave in a manner that is not representative of their behavior outside the research setting. In addition, interpretation of the study's findings can be threatened because participants may unintentionally make a research variable more effective than it actually is, or even nullify the effect of an otherwise important variable. One way to reduce the problem of demand characteristics is to limit participants' knowledge about their role in the study or about the hypotheses of the study, that is, to provide as few "cues" as possible. You may remember, however, that withholding information from participants can raise ethical concerns, particularly concerning informed consent.

Key Concept

Controlling Reactivity There are several approaches that researchers use to control the problem of reactivity. Several of the observational methods discussed earlier in this chapter are designed to limit reactivity. Reactivity can be eliminated if research participants do not know that an observer is present in the setting. Disguised participant observation achieves this goal because individuals are not aware of the presence of the observer. We can presume, then, that they behave as they normally would. Recall that this procedure was used in Rosenhan's (1973) study of hospitalization of the mentally ill and social psychologists' observations of individuals who claimed to be in contact with aliens

(Festinger et al., 1956). Observers can also conceal themselves while making observations in natural settings (naturalistic observation), as seen in the stop-sign study, or they can use hidden cameras or tape recorders to make their observations (but they must be aware of ethical concerns related to privacy).

An important advantage of indirect observation, or unobtrusive methods, is that these observations are nonreactive. Researchers observe physical traces and archival records to learn about people's past behavior. Because the individuals are no longer present in the situation and likely do not even know the physical traces or archival records are being observed by researchers, it is impossible for them to change their behavior. One researcher investigated the drinking behavior of people living in a town that was officially "dry" by counting empty liquor bottles in their trash cans (see Figure 4.6). Another researcher used the archival records kept by a library to assess the effect of the introduction of television in a community. Withdrawal of fiction titles dropped, but the demand for nonfiction was not affected (see Webb et al., 1981). It would be interesting to conduct a similar study today, considering the widespread availability of science, history, and biography cable programs. One might hypothesize that the advent of these programs is correlated with a decline in nonfiction rentals from libraries.

Another approach researchers use to deal with reactivity is to adapt participants to the presence of an observer. We can assume that as participants get used to an observer's presence, they will eventually behave normally in the observer's presence. Adaptation can be accomplished through either habituation or desensitization. In a *habituation* procedure, observers simply enter

FIGURE 4.6 Unobtrusive (nonreactive) measures of people's behavior can be obtained by searching their trash for physical traces, but ethical issues regarding privacy must be considered.

into the setting on many different occasions until the participants stop reacting to their presence (i.e., their presence becomes normal). Habituation was used to film a documentary titled *An American Family*, which was shown on public television in the 1970s. The camera crew literally moved into a California home and recorded the family for seven months. Although it's impossible to tell how much the family's behavior was influenced by the presence of these observers, the events that unfolded on camera provided evidence that family members had habituated to the cameras. Most notably, the family broke up and the wife asked the husband to move out of the house. When interviewed later about having the divorce announced to millions of viewers, the husband said that although they could have asked the camera crew to leave, by that time, he said, "we had gotten used to it" (*Newsweek*, 1973, p. 49). It's likely that similar processes of habituation take place during more contemporary "reality shows," but one must also wonder whether some of the behavior displayed on these shows occurs precisely *because* the individuals are on television!

Desensitization as a means of dealing with reactivity is similar to the procedures used by clinical psychologists in the behavioral treatment of phobias. In a therapy situation, an individual with a specific fear (e.g., spiders) is first exposed to the feared stimulus at a very low intensity. For example, the individual may be asked to think of things related to spiders, such as cobwebs. At the same time, the therapist helps the client to practice relaxation. Gradually the intensity of the stimulus is increased until the client can tolerate the actual feared object, for example, by holding a spider. Desensitization is often used by animal researchers to adapt animal subjects to the presence of an observer. Prior to her violent murder in Africa, Dian Fossey (1981, 1983) conducted fascinating observational studies of mountain gorillas in Rwanda. Over a period of time she moved closer and closer to the gorillas so they could adapt to her presence. She found that by imitating their movements—for instance, by munching the foliage they ate and by scratching herself—she could put the gorillas at ease. Eventually she was able to sit among the gorillas and observe them as they touched her and explored her research equipment.

Ethical Issues Whenever researchers try to control reactivity by observing individuals without their knowledge, important ethical issues arise. For instance, observing people without their consent can represent a serious invasion of privacy. Deciding what constitutes an invasion of privacy is not always easy (see Chapter 3), and must include consideration of the sensitivity of the information, the setting where observation takes place, and the method for disseminating the information (e.g., Diener & Crandall, 1978).

Recent behavioral studies using the Internet introduce new ethical dilemmas. For example, when researchers enter Internet chat rooms as disguised participant observers to find out what makes racists advocate racial violence (Glaser et al., 2002), the information they obtained could be seen as incriminating evidence without the respondents' knowledge, much like a "sting" operation. The dilemma, of course, is that if informed consent were obtained it is very unlikely that respondents would cooperate. In this case, the IRB approved the research by agreeing with the researchers that a chat room is a "public forum," that these

topics were common to that forum, and that the researchers had appropriately established safeguards to protect respondents' identities (e.g., by separating names or pseudonyms from comments). On the other hand, there are instances in which people have felt their privacy was violated when they learned that researchers observed their online discussions without their knowledge (see Skitka & Sargis, 2005). Although Internet message boards may be considered "public," researchers investigating adolescent messages about self-injurious behaviors were required by their university IRB to paraphrase participants' comments rather than use exact quotes (Whitlock et al., 2006). Behavioral research using the Internet is in its early stages, and both researchers and IRB members are still learning and applying creative problem solving for ethical dilemmas as they arise (see Kraut et al., 2004).

When individuals are involved in situations that are deliberately arranged by an investigator, as occurs in structured observation and field experiments, ethical problems associated with placing participants at risk may arise. Consider, for example, a field experiment in which students walking across campus were questioned about their attitudes toward racial harassment (Blanchard, Crandall, Brigham, & Vaughn, 1994). In one condition of the experiment, a confederate, posing as a student, condemned racist acts and in a second condition, the confederate condoned racist acts. Individual participants were then asked about their attitudes. The results of the study indicated that the views expressed by the confederate caused participants to be more likely to express similar statements compared to a third condition, in which the confederate didn't express any opinion. We can ask, were these participants "at risk"? Did the goals of the study, which were to show how outspoken people can influence interracial social settings, outweigh any risks involved in the study? Although participants were "debriefed immediately" in this study, is that sufficient to address any concerns about how they may have behaved when confronted with racist opinions? Did debriefing restore their confidence in a science that seeks knowledge through deception? Any attempt to answer these questions highlights the difficulty of ethical decision making.

Finally, we can turn to unobtrusive measures such as physical traces and archival data to address another ethical issue: scientists' ethical obligation to improve individual and societal conditions. There are many serious issues that confront us today, including violence, race relations, suicide, domestic conflict, and many other social issues, for which research involving direct observation may be difficult to justify when considering a risk/benefit ratio. That is, some research methods simply may involve too great a risk to research participants. However, psychologists' ethical obligation to improve the conditions of individuals, organizations, and society requires that they seek methods to gain knowledge in these important areas, for the cost of *not* doing research to solve these problems is high. Research involving the use of physical traces and archival data can be carried out on these important problems under conditions where ethical issues are often minimal relative to more intrusive methods. Thus, unobtrusive observational methods represent an important tool in the multimethod approach for investigating important social issues with less risk.

Observer Bias

- Observer bias occurs when researchers' biases determine which behaviors they choose to observe, and when observers' expectations about behavior lead to systematic errors in identifying and recording behavior.
- Expectancy effects can occur when observers are aware of hypotheses for the outcome of a study or the outcome of previous studies.
- The first step in controlling observer bias is to recognize that it may be present.
- Observer bias may be reduced by keeping observers unaware ("blind") of the goals and hypotheses of the study.

As an example of disguised participant observation, we described Rosenhan's (1973) classic study in which observers were admitted to psychiatric hospitals. Once in the hospital they observed and recorded behavior of hospital staff. Rosenhan's research identified a serious bias on the part of the staff. Once the observers (called "pseudopatients") were labeled schizophrenic, staff members interpreted their behavior solely according to this label. Behaviors that otherwise might be considered normal were interpreted by the staff as evidence of the pseudopatients' illness. For instance, the pseudopatients quickly learned they could record their observations openly—no one paid much attention to what they were doing. When Rosenhan later checked the medical records for the pseudopatients, he found that staff members had cited the note taking as a symptom of their illness. (Don't worry—taking notes is not a sign of mental illness!) Because staff members interpreted the pseudopatients' behavior in terms of the schizophrenic label, their "sanity" was not detected. This example clearly illustrates the danger of **observer bias,** the systematic errors in observation that result from an observer's expectations. In this case, the staff members demonstrated observer bias.

Key Concept

Expectancy Effects In many scientific studies the observer has some expectations about what behavior should be like in a particular situation or following a specific psychological treatment. When researchers design a study they review the previously published research literature to help them develop their hypotheses. This knowledge can lead researchers to form expectancies about what should occur in a research situation; in fact, hypotheses are predictions about what is expected to happen. However, expectancies can be a source of observer bias—*expectancy effects*—if they lead to systematic errors in observation (Rosenthal, 1966, 1976). A classic study documented expectancy effects (Cordaro & Ison, 1963). College student observers recorded the number of head turns and body contractions made by flatworms. Observers in one group were led to expect a high rate of movement, whereas observers in a second group expected a low rate. The two groups of flatworms were essentially identical; however, results showed that when observers expected to see lots of movement, they recorded twice as many head turns and three times as many body contractions compared to observers who expected a low rate of movement. Apparently, the students systematically interpreted the actions of the worms differently depending on what they expected to observe.

Other Biases An observer's expectancies regarding the outcome of a study may not be the only source of observer bias. You might think that using automated equipment such as video cameras would eliminate observer bias. Although automation reduces the opportunity for bias, it does not necessarily eliminate it. Consider the fact that, in order to record behavior on film, the researcher must determine the angle, location, and time of filming. To the extent that these aspects of the study are influenced by the researcher's personal biases, such decisions can introduce systematic errors into the results. For example, Altmann (1974) described an observational study of animal behavior in which the observers biased the results by taking midday breaks, which coincided with a time of relative inactivity among the animals. Observations of the animals during these periods of inactivity were conspicuously absent for the observational records, which biased the results to make the animals appear more active than they were. In addition, the use of automated equipment generally only postpones the process of classification and interpretation, and it is perfectly possible for the effects of observer bias to be introduced when narrative records are coded and analyzed.

Controlling Observer Bias Observer bias is difficult to eliminate, but it can be reduced in several ways. As we mentioned, the use of automatic recording equipment can help, although the potential for bias is still present. *Probably the most important factor in dealing with observer bias is the awareness that it might be present.* That is, an observer who knows about this bias will be more likely to take steps to reduce its effect. One important way researchers reduce observer bias is to limit the information provided to observers. When observers and coders do not know the hypotheses of a study they cannot form expectations about behavior. In a manner of speaking, observers can be kept "blind" regarding certain aspects of the study. Observers are blind when they do not know the reasons for the observations or the goals of the study. For example, when trained coders analyzed the videotapes of interactions between mothers and children from maltreating and nonmaltreating families, they were not aware of which type of family they were observing (Valentino et al., 2006). As you might imagine, observers' expectancies regarding maltreating families might influence their interpretation of behaviors, just as staff members in Rosenhan's (1973) study interpreted pseudopatients' behavior according to their diagnostic label. Using blind observers greatly reduces the possibility of introducing systematic errors due to observer expectancies.

Summary

Researchers rarely observe all behavior that occurs. Consequently, researchers must use some form of behavior sampling such as time and situation sampling. An important goal of sampling is to achieve a representative sample of behavior. External validity refers to the extent to which observations from a study can be generalized to describe different populations, settings, and conditions; external validity is enhanced when a representative sample is obtained. Observational methods can be classified as direct observation or

indirect observation. Direct observation in a natural setting without intervention is called naturalistic observation. Observation with intervention can take the form of participant observation, structured observation, and field experiments. An important advantage of indirect observational methods is that they are nonreactive. Reactivity occurs when people change their behavior because they know they are being observed. Indirect, or unobtrusive, observations can be obtained by examining physical traces and archival records. Physical traces include use traces (natural or controlled) and products. Archival data are the records of the activities of individuals, institutions, governments, and other groups. Problems associated with physical traces include potential biases in how traces accumulate or survive over time, and problems with archival data include selective deposit, selective survival, and the potential for spurious relationships in the data.

In observational studies, behavior can be recorded either with a comprehensive description of behavior or by recording only certain predefined units of behavior. Narrative records are used to provide comprehensive descriptions of behavior, and checklists typically are used when researchers are interested in whether a specific behavior has occurred (and under what conditions). Frequency, duration, and ratings of behavior are common variables examined in observational studies. The analysis of narrative records involves coding as one step in data reduction. Content analysis is used to examine archival records. How quantitative data are analyzed depends on the measurement scale used. The four measurement scales are nominal, ordinal, interval, and ratio. When a nominal scale is used to record behavior (e.g., present, absent), data are summarized using proportions or percentages to indicate relative frequency of behavior. When describing ordinal data, researchers often describe results according to the percentage of people who ranked items first among a set of items. When behavior is measured using interval and ratio scales, data are summarized using the mean and standard deviation. It is essential to provide measures of observer reliability when reporting the results of an observational study. Depending on the level of measurement used, either a percentage agreement measure or a correlation coefficient can be used to assess reliability.

Possible problems due to reactivity or observer bias must be controlled in any observational study. One form of reactivity is when participants pay attention to the demand characteristics of a research situation to guide their behavior. Observational methods in which the participants are not aware they are being observed (e.g., disguised participant observation, unobtrusive methods) limit reactivity; in other situations, participants may adapt to the presence of an observer. Observer bias occurs when researchers' biases determine which behaviors they choose to observe and when observers' expectations about behavior lead to systematic errors in identifying and recording behavior (expectancy effects). Important steps in reducing observer bias are to be aware of its presence and to keep observers blind regarding the goals and hypotheses of the study. Ethical issues must be considered prior to beginning any observational study. Depending on the nature of the observations, ethical issues might include deception, privacy, informed consent, and the risk/benefit ratio.

Key Concepts

external validity 94
time sampling 94
situation sampling 96
naturalistic observation 97
participant observation 100
reactivity 100
structured observation 103
field experiment 105
unobtrusive measures 107
physical traces 107
archival records 110
selective deposit 111
selective survival 112
narrative records 113
measurement scales 115
data reduction 120
coding 120
content analysis 120
interobserver reliability 124
correlation coefficient 125
demand characteristics 127
observer bias 131

Review Questions

1. Describe the types of sampling researchers use in observational studies and what the proper use of sampling is intended to accomplish.
2. Explain the difference between direct and indirect observational methods and how the degree of intervention can be used to distinguish direct observational methods.
3. Describe a research situation in which naturalistic observation can be useful when ethical considerations prevent researchers from intervening to study behavior.
4. Explain why reactivity is a problem in observational studies.
5. Explain how structured observation represents a "middle ground" in psychological research and identify the primary advantage and potential cost of this compromise.
6. Explain why physical traces and archival data are attractive alternatives to direct observation.
7. Describe the different types of physical-trace measures available to psychologists and the ways in which these measures may be biased.
8. Explain how archival data may be used to test the effect of a natural treatment.
9. Explain how selective deposit, selective survival, and spurious relationships may bias the interpretation of archival records.
10. Describe how data reduction and coding are used in qualitative analyses of narrative records and archival data.
11. Give an example using each of the four measurement scales to describe how a researcher could measure eye contact between pairs of people in conversation with each other.
12. What are the most common descriptive measures (a) when events are measured on a nominal scale, (b) when items are ranked using an ordinal scale, and (c) when behavior is recorded on at least an interval scale.
13. Describe the procedures researchers can use to increase interobserver reliability.
14. Identify the measurement scales that require a correlation coefficient to assess interobserver reliability, and explain what a negative correlation would indicate in this situation.
15. Explain whether high interobserver reliability ensures that the observations are accurate and valid.
16. Describe two ways in which observer bias (expectancy effects) can occur in psychological research.
17. Explain how researchers may reduce observer bias.

CHALLENGE QUESTIONS

1 Students in a developmental psychology lab course conducted an observational study of parent–infant interactions in the home. When they first entered the home on each of the 4 days they observed a given family, they greeted both the parents and the infant (and any other children at home). They instructed the family to follow its daily routine, and they asked a series of questions about the activities of that day to determine whether it was a "normal" day or whether anything unusual had happened. The students tried to make the family feel comfortable, but they also tried to minimize their interactions with the family and with each other. For any given 2-hour observation period there were always two student observers present in the home, and the two observers recorded their notes independently of each other. Each of six pairs of students was randomly assigned to observe two of the 12 families who volunteered to serve in the study. The same pair of observers always observed a given family for the entire 8 hours of observation for that family. The observers used rating scales to record behaviors on a number of different dimensions, such as mutual warmth and affection of the parent–infant interaction.

- **A** Cite two specific procedures used by the students to ensure the reliability of their findings.
- **B** Cite one possible threat to the external validity of the findings of this study; once again, cite a specific example from the description provided.
- **C** Cite one specific aspect of their procedure that indicated that the students were sensitive to the possibility that their measurements might be reactive. What other methods might they have used to deal with this problem of reactivity?

2 An observational study was done to assess the effects of environmental influences on drinking by college students in a university-sponsored pub. Eighty-two students over the age of 21 were observed. The observers used a checklist to record whether the participant was male or female and whether the participant was with one other person or was in a group of two or more other people. Each observation session was always from 3 P.M. to 1 A.M., and observations were made Monday through Saturday. The observations were made over a 3-month period. Two observers were always present during any observation session. Each participant was observed for up to 1 hour from the time he or she ordered the first beer. The data were summarized in terms of the number of beers drunk per hour. The results showed that men drank more and men drank faster than did women. Men drank faster when with other men, and women also drank faster with men present. Both men and women drank more in groups than when with one other person. These results indicate that the environment within which drinking occurs plays an important role in the nature and extent of that drinking.

- **A** Identify the observational method being used in this study, and explain why you decided on the observational method you chose.
- **B** Identify the independent and dependent variables in this study, and describe the operational definition of each level of the independent variable.
- **C** How could the researchers control for reactivity in this study? What ethical concerns might arise from their approach?
- **D** Identify one aspect of the procedures in this study that would likely *increase* the reliability of the observations.
- **E** Identify one aspect of the procedures in this study that would likely *limit* the external validity of the findings of this study.

3 A bright female graduate student in psychology has been offered a job with both *Newsweek* and *Time*. The salary offers of the two companies are basically the same, and it appears that both the working conditions and the job responsibilities are similar. To help her decide which job to accept, she resolves to determine whether one magazine has a better attitude toward women than the other. She appeals to you to help her with a content analysis of these two news magazines. What specific advice would you give her regarding each of the following steps of her content analysis?

- **A** Sampling
- **B** Coding
- **C** Reliability
- **D** Quantitative and qualitative measures

4 Four students were doing internships at the Social Science Research Institute of their university. The research institute had a contract to do a series of studies on traffic safety for the downtown development agency of a small city near the university. The internship students were assigned to carry out one of the studies. Specifically, they

(*continued*)

were to do a study to determine how likely it was that cars actually came to a stop at intersections with stop signs with pedestrian crosswalks in the downtown area. You are to respond to the following questions that the students are considering in planning their study.

A The students want to distinguish the extent to which the cars stop beyond a "yes" or "no" classification. How could the students develop an operational definition for the cars stopping that would include cars that came to a full stop, came to a rolling stop, and did not stop at all?

B What steps could the students take before beginning to collect data for the actual study to increase the interobserver reliability of their observations?

C The students are interested in determining the likelihood that cars will stop when pedestrian traffic downtown is light and when it is heavy. What time-sampling plan could the students use to make this determination?

D The students are especially interested in determining the likelihood of cars stopping at the stop sign independent of whether other cars have stopped. How would the students need to sample the cars they observed in order to study the independent stopping of cars? What information could the students record that would allow them to include all cars in their sample and still determine the likelihood of cars stopping independently?

Answer to Stretching Exercise

1 Because the students did not intervene in the situations (natural settings) they observed, this study is best described as naturalistic observation.

2 The students' choice of a 5-minute observation interval may have limited their ability to measure concentration effectively. The time interval may have been too short to show changes in concentration, making it difficult to detect differences between the two locations.

3 Time-sampling is important in this study because students' ability to concentrate may vary across days of the week and times of the day. By choosing only one time period (Monday, 9 to 11 P.M.), the external validity of the study is limited. Sampling different times of the day, days of the week, and weeks of the semester would improve the external validity of the study.

4 One possibility is that students choose different types of material to study in the two locations. If studying in a student union is more difficult, then students may choose easier material that requires less effort to maintain concentration while studying in the student union. This difference in study material might account for the observation that concentration times did not differ. One of the challenges of naturalistic observation is that researchers cannot control factors (e.g., type of material studied) that could influence the outcome of observations.

Answer to Challenge Question 1

A The students' procedures that enhanced reliability were as follows: observing each family for 8 hours, using two independent observers, and using checklists to provide operational definitions.

B One possible threat to the external validity of the findings was that the 12 families volunteered for the study and such families may differ from typical families.

C The students' efforts to minimize interactions with the family and with each other suggests that they were sensitive to the problem of reactivity. Two other methods they might have used are habituation and desensitization.