PART TWO

Descriptive Methods

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 behavior—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 be 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. However, our everyday observations and those of a scientist differ in many ways. When we observe casually, we may not be aware of factors that bias our observations, and we rarely keep formal records of our observations. Instead, we rely on our memory of events even though our own experience (and the observations of psychologists) confirms that our memory is not perfect!

Scientific observation, on the other hand, is made under precisely defined conditions, in a systematic and objective manner, and with careful record keeping. Matsumoto and Willingham (2006), for example, recorded the facial behaviors of medal winners in the judo competition at the 2004 Athens Olympic games. A professional photographer took thousands of pictures during the 7-day competition. The researchers focused on the facial expressions of bronze and gold medal winners during the seconds following their winning the medal match. These spontaneous expressions were recorded from athletes from 35 different countries, providing a highly diverse sample of people experiencing a similar emotionally charged event. Examination of the photographs was done systematically using a highly sophisticated coding system that identified each of the functionally anatomical facial muscle movements. Results of this investigation provided evidence for the universal nature of the expressed emotions and for their reliable signaling of the prior event.

Systematic observation is an important tool not only of psychologists but of anthropologists, sociologists, and ethologists as well. In this chapter we focus on observation in natural settings, but you should remember that the principles of good observation apply equally well in natural settings and in the laboratory.

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 context in which the behavior occurs. Consider your own behavior in this regard. Do you behave the same at home as in school, or at a party compared to in a classroom? Does your observation of others 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. For example, Kagan, Reznick, and Snidman (1988) observed the reactions of a group of 2- and 3-year-old children in the presence of an unfamiliar person or object. The researchers identified those children who were consistently shy, quiet, and timid and those who were consistently sociable, talkative, and affectively spontaneous. When the same children were observed at 7 years of age, a majority of the children in each group exhibited similar behaviors. Based on these observations, the researchers developed a hypothesis relating shyness in childhood to extreme social anxiety in adulthood. They argued that both were the result of temperamental differences present at birth. Similarly, Caspi and his colleagues (1997) found that observations made at age 3 could be used to successfully predict health-risk behaviors (e.g., abusing alcohol) of young adults. These findings support the hypothesis that temperamental differences observed early in childhood are linked to distinct adult behavior patterns.

In this chapter you will see that the scientist-observer is not always passively recording behavior as it occurs. We take a look at reasons why scientists intervene to create special situations for their observations. 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 studies using observation.

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, researchers can *generalize* their findings to that population. The sampling of times, situations, and conditions strongly influences the most important dimension of sampling—who the participants will be. 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 to describe people, settings, and conditions beyond those used in the study. In this section we describe how time sampling 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. 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. Assume that researchers in this study want to observe the children for a total of 2 hours each day. 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 beginning every 2 hours. The first observation period could begin at 9 A.M., the second at 11 A.M., and so forth. Another possibility would be to schedule 10-minute observation periods every half hour during the school day. A *random* time-sampling technique could be used in the same situation by distributing four 30-minute periods (or a dozen 10-minute periods) randomly over the course of the day. A different random schedule would be determined each day on which 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 participants 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, having scheduled four 30-minute observation periods at the same time each day (e.g., 9 A.M., 11 A.M., etc.), a researcher might then decide to observe only during 20-second intervals that are randomly

distributed within each 30-minute period. Whatever time-sampling procedure is used, the observer must carefully consider both the advantages and limitations of the schedule in terms of its potential to yield a representative sample of behavior.

Time sampling is not an effective method of 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. Event sampling is a more effective and efficient sampling method for infrequent events. 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. For example, in a study of children's "rough-and-tumble" play, an observer positioned herself in the corner of a playground to observe members of a nursery school class (Smith & Lewis, 1985). Due to the relatively low frequency of this behavior, event sampling was the method of choice in this study. The researcher made observations whenever rough play began and continued to observe until the given episode of rough play ended. Although event sampling is an efficient method for observing unpredictable events, the use of event sampling can easily introduce biases into the behavioral record. 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. In most situations, an observer is likely to achieve a representative sample of behavior only when some form of time sampling is used.

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 different situations, researchers reduce the chance that their results will be peculiar to a certain set of 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. Mutual gaze between a mother and an infant occurs in chimpanzees 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 than could be claimed if only particular types of individuals were observed. For example, as part of a naturalistic observation of beer drinking among college students, investigators purposely sampled behavior in various settings where beer was served, including five town bars and a student center (see Geller, Russ, & Altomari, 1986). In a different study, 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 class, 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 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 (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 at the dining hall.

Classification of Observational Methods

- Observational methods can be classified as "observation with intervention" or "observation without intervention."
- Methods for recording behavior can be classified in terms of how much of the behavior is recorded.

Observational methods can be classified on two dimensions (Willems, 1969). The first important distinction is between observation with intervention and observation without intervention. The second dimension involves the methods of recording behavior. Observation studies can be distinguished in terms of whether all (or nearly all) of the behavior is recorded or whether only particular units of behavior are recorded. In some situations, researchers seek a comprehensive description of behavior. They accomplish this by recording behavior using film, tapes, or lengthy verbal descriptions. More often, researchers record specific units of behavior that are related to the goals of a particular study. For example, Chambers and Ascione (1987) observed children playing either an aggressive video game or a video game with prosocial content. Children who played the aggressive game were later observed to put less money in a donation box and were less likely to help sharpen pencils than children who played the prosocial game. Thus, in this study, the researchers recorded specific responses

related to the children's prosocial behavior. Similarly, researchers interested in public smoking among youth focused solely on the smoking behaviors of adults and minors in several Illinois towns (Jason, Pokorny, Sanem, & Adams, 2006).

We will discuss observational methods first in terms of the extent of observer intervention and then in terms of methods for recording behavior.

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

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 what occurs. The events occur naturally and are not manipulated or controlled by the observer. 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.

Observing people in a psychology laboratory would not be considered naturalistic observation. The laboratory situation has been created specifically to study behavior. In that sense, the laboratory is an artificial rather than a natural setting. In fact, observation in natural settings serves, among other functions, as a way of establishing the external validity of laboratory findings—bringing the lab into the "real world." 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 oriented) 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.

FIGURE 4.1 Animal researchers such as Jane Goodall frequently rely on naturalistic observation to obtain information about the behavior of their subjects.

Hartup's study of children's aggression illustrates why a researcher may choose to use naturalistic observation rather than to manipulate experimental conditions related to 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 this problem 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); descriptions 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 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.

Psychologists are not the only researchers who observe behavior in natural settings. Observation is a fundamental method in ethology. Although related

to psychology, *ethology* is generally considered a branch of biology (Eibl-Eibesfeldt, 1975). Ethologists study the behavior of organisms in relation to their natural environment. They adopt a comparative approach to understanding behavior and frequently seek to explain behavior in one animal species on the basis of innate patterns of behavior observed in species lower on the evolutionary scale. Speculations about the role of innate mechanisms in determining human behavior are not uncommon among ethologists. The focus of an ethological investigation is often the development of an *ethogram.* This is a complete catalog of all the behavior patterns of an organism, including information on frequency, duration, and context of occurrence for each behavior.

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.

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. Although the types of intervention employed by psychologists are too numerous and diverse to classify, their reasons for intervening are generally one or more of the following:

1 To precipitate or cause an event that occurs infrequently in nature or that normally occurs under conditions that make it difficult to observe.
2 To investigate the limits of an organism's response by varying systematically the qualities of stimulus event.
3 To gain access to a situation or event that is generally not open to scientific observation.
4 To arrange conditions so that important antecedent events are controlled and consequent behaviors can be readily observed.
5 To establish a comparison by manipulating one or more independent variables to determine their effect on behavior.

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.

Participant Observation

- Undisguised participant observation is often used to understand the culture and behavior of groups of individuals.
- Disguised participant observation is often used when researchers believe individuals would change their behavior if they knew it was being recorded.
- Participant observation allows researchers to observe behaviors and situations that are not usually open to scientific observation.

- Participant observers may sometimes lose their objectivity or may unduly influence the individuals whose behavior they are recording.

Key Concept

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. Politicians, for instance, often make different statements when speaking to the press, depending on whether their comments are "for" or "off" the record. Our own behavior is likely to be affected by knowing that we are being watched. Because of this possibility, researchers may decide to disguise their role as observers if they believe that people being observed will not act as they ordinarily would if 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.

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 participants: Eight individuals (including psychologists, a pediatrician, and a housewife) misrepresented their names, occupations, and symptoms and sought admission to 12 different 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 researchers stopped complaining of any symptoms and refrained from acting abnormally (except, for a while, their continued anxiety about being "caught"). In addition to observing patient–staff interactions, the observers were interested in how long it took for a "sane" person to be released from the hospital. The researchers were hospitalized from 7 to 52 days, and when they were discharged, their schizophrenia was said to be "in remission." Apparently, once the pseudopatients were labeled schizophrenic, they were stuck with that label through their discharge. There are, however, reasons to challenge this specific conclusion and other aspects of Rosenhan's (1973) study (see Box 4.1).

A participant observer may also be in a position to have the same experiences as the people under study. This experience may provide important insights and understanding of individuals or groups. The pseudopatients in the Rosenhan

BOX 4.1

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.

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 the objectivity required for valid observations if they identify with the individuals under study. Changes in a participant observer are sometimes dramatic and are not easily anticipated. Witness the experiences of a criminologist who used

undisguised participant observation to study police officers at work. Kirkham (1975) went through police academy training like any recruit and became a uniformed patrol officer assigned to a high-crime area in a city of about half a million. His immersion in the daily activities of an officer on the beat led to marked changes in his attitudes and personality. As Kirkham himself noted:

> As the weeks and months of my new career as a slum policeman went by, I slowly but inexorably began to become indistinguishable in attitudes and behavior from the policemen with whom I worked. . . . According to the accounts of my family, colleagues and friends, I began to increasingly display attitudinal and behavioral elements that were entirely foreign to my previous personality—punitiveness, pervasive cynicism and mistrust of others, chronic irritability and free-floating hostility, racism, a diffuse personal anxiety over the menace of crime and criminals that seemed at times to border on the obsessive. A former opponent of capital punishment, I became its vociferous advocate in cases involving felony murder, kidnapping and the homicide of police officers—even though as a criminologist I continued to recognize its ineffectiveness as a deterrent to crime. (p. 19)

Participant observers must be aware of the threat to objective reporting that arises due to their involvement in the situation they are studying. This threat necessarily increases as degree of involvement increases.

Another potential problem with observer involvement is the effect the observer can have on the behavior of those being studied. It is more than 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. Whenever observers intervene in a natural setting, they must ask to what degree participants and events are affected by their intervention. Is what is being observed the same as it would have been if the observer had never appeared? It is difficult to generalize results to other situations if intervention produces behavior that is specific to the conditions and events created by the observer.

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, and the role of the observer in the group. The disguised participant observation in the Rosenhan study appears to have been successful. Rosenhan and his associates seem not to have significantly affected the natural environment of the hospital unit by assuming the role of patients. However, some of the patients—though none of the staff—apparently detected the sanity of the pseudopatients, suggesting to the observers that they were there to check up on the hospital.

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 participants' 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). A leader of the group said he had received a message from the aliens predicting a cataclysmic flood on a specific date. The flood was to stretch from the Arctic Circle to the Gulf of Mexico. 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 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, the flood never occurred, but at least some of the group members came to use this disconfirmation as a means of strengthening their initial belief. They began to seek new members by arguing that 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 that a participant observer may have on the behavior under study.

Structured Observation

- Structured observations are set up to record behaviors that may be difficult to observe using naturalistic observation.
- Structured observations are often used by clinical and developmental psychologists.
- Problems in interpreting structured observations can occur when the same observation procedures are not followed across observations or observers, or when important variables are not controlled.

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 7). 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 than they would be without intervention. For example, in order to study children's understanding of intentional acts, a female adult sometimes handed infants a toy. Other times the adult was unwilling to give it away (e.g., she played with it herself) or was unable to give it away because she "accidentally" dropped it. The researchers observed that infants at 9 months of age reacted with more impatience when the adult was unwilling to give them the toy than when it was accidentally dropped; younger infants showed no such differentiation (Behne, Carpenter, Call, & Tomasello, 2005).

Key Concept

In other types of structured observation, researchers may create quite elaborate procedures to investigate a particular behavior more fully. Simons and Levin (1998) used structured observation to study a phenomenon called change blindness. Change blindness occurs when people fail to notice changes in their environment. With so much going on around us, it is impossible for us to notice every little change. Simons and Levin demonstrated, however, that people often fail to notice changes even when they are paying attention. In their study the researchers used *confederates,* that is, individuals in the research situation who are instructed to behave in a certain way in order to create a situation for observing

FIGURE 4.2 Frames from a video of a subject from Experiment 1. Frames *a–c* show the sequence of the switch. Frame *d* shows the two experimenters side by side.

a b c d

behavior. One confederate approached a pedestrian walking across campus and asked for directions. About 15 seconds into the conversation, two other confederates rudely passed between them carrying a door. As the door passed, the original confederate and one of the confederates carrying the door changed places. This structured observation created a changed environment with the new confederate now conversing with the pedestrian (see Figure 4.2). The new confederate typically made eye contact with the pedestrian and differed from the original confederate in height, voice, and clothing. Only about half of the pedestrians noticed the switch; half of the pedestrians were blind to the change.

Structured observations may occur in a natural setting as in the Simons and Levin (1998) study or in a laboratory setting. Psychologists often use structured observations when making behavioral assessments of parent–child interactions. For example, researchers have observed play between mothers

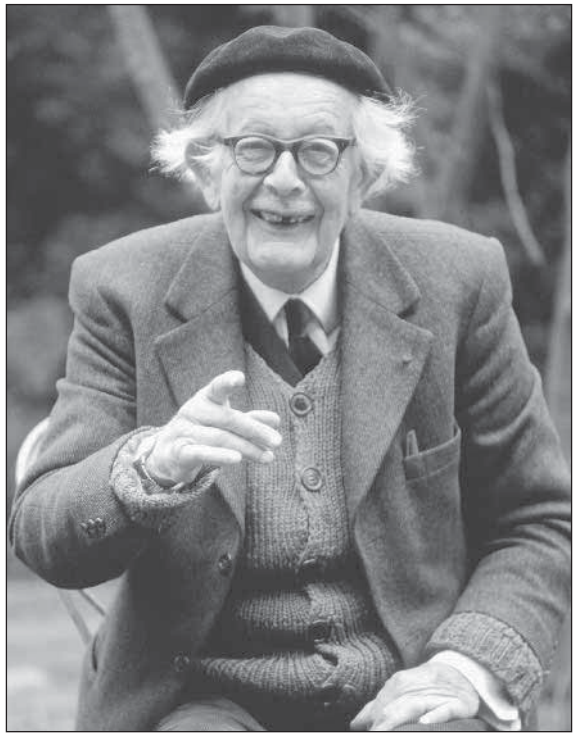
and children from maltreating (e.g., abusing, neglecting) families and nonmaltreating families (Valentino, Cicchetti, Toth, & Rogosch, 2006). Mothers were videotaped in a laboratory setting through a one-way mirror while interacting with their children in different contexts arranged by the researchers. The researchers observed that children from abusing families engaged in less independent play than children from nonmaltreating families and mothers from 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 early intervention.

Animal researchers also frequently structure an observational situation to create events that would normally be difficult to observe in natural settings. An interesting example in this regard comes from a series of studies investigating differences in the cognitive capacity of apes and humans. (Was Darwin right? Is it just a matter of degree?) One type of structured observation involves assessing the gaze-following behavior of chimpanzees (see Povinelli & Bering, 2002). A human caretaker makes eye contact with the chimpanzee and then looks above the chimpanzee. Observations reveal that chimpanzees do indeed follow the caretaker's gaze (i.e., looking above); however, controversy exists as to what this behavior really means. Do apes "understand" the mental state of the human caretaker (i.e., that the person has seen something)?

Developmental psychologists frequently use structured observations. Jean Piaget (1896–1980) is perhaps most notable for his use of these methods. In many of Piaget's studies, a child is first given a problem to solve and then given several variations of the problem to test the limits of the child's understanding. The observer acquaints the child with the nature of the problem and then asks questions to probe the child's reasoning processes. 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). Piaget-type tasks are also frequently used by animal researchers to investigate aspects of animal cognition. In one such study, orangutans outperformed squirrel monkeys on a classic test of object permanence, the ability to represent cognitively the existence of an unperceived (e.g., unseen) object (de Blois, Novak, & Bond, 1998).

Structured observation represents a compromise between the passive nonintervention of naturalistic observation and the systematic manipulation of independent variables and precise control that characterize laboratory experiments. The advantage of such a compromise is that it permits observations to be made under conditions that are more natural than the artificial conditions imposed in a laboratory. Nevertheless, there may be a price to pay. The failure to follow similar procedures each time an observation is made may make it 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. For example, observers who use structured observations like those used by Piaget do not always follow the same procedure from one observation to another. This inconsistency in procedure across observations is a potential problem with these techniques (see Brainerd, 1978).

FIGURE 4.3 Jean Piaget (1896–1980) used structured observation to investigate children's cognitive development.



Field Experiments

- In a field experiment, researchers manipulate one or more independent variables in a natural setting to determine the effect on behavior.

Key Concept

When an observer manipulates one or more independent variables in a natural setting in order to determine their effect on behavior, the procedure is called a **field experiment.** We mention field experiments in this chapter because they represent 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. In a field experiment, researchers typically manipulate an independent variable to create two or more conditions and they measure the effect of the independent variable on behavior. Field experiments are frequently used in social psychology (Bickman, 1976). For example, confederates have posed as robbers when bystander reaction to a crime has been investigated (Latané & Darley, 1970) and as individuals cutting into a waiting line in order to study the reactions of those already in line (Milgram, Liberty, Toledo, & Wackenhut, 1986). Our discussion of experimental methods will continue in Chapter 7.

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. Do you think the operational definition of concentration used for the dependent variable "captured" students' concentration? How could their operational definition of the dependent variable have contributed to both the high overall level of study time they observed and the lack of a difference between the two locations? What would you recommend to improve the operational definition of concentration in this study?
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?

RECORDING BEHAVIOR

- The goals of the 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.

Observational methods also differ in the manner in which behavior is recorded. A particularly important feature is the degree to which behavior is abstracted from the situation in which it is observed. Whether all behavior or only *selected* aspects are observed in a setting depends on the purpose of the study and the researchers' goals. Although researchers sometimes seek a *comprehensive* description of behavior and the situation in which it occurs, more often they focus on only certain behaviors or events. Decisions regarding

how behavior is recorded also depend on whether the investigator is doing qualitative or quantitative research. The results of a qualitative study are presented chiefly in the form of verbal description and logical argument. Reports of quantitative research mainly emphasize statistical description and analysis of data to support a study's conclusions. *The most important point to remember is that how you choose to record behavior determines how the results of your study are eventually measured, summarized, analyzed, and reported.*

Qualitative Records of Behavior

- Narrative records in the form of written descriptions of behavior, and audio and video recordings, are comprehensive records of observed behavior.
- 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 recordings, videos, and films. We mentioned, for example, that videos were used to capture the mother–child interactions among maltreating and nonmaltreating families (Valentino et al., 2006). Because ethologists typically are interested in every detail of a behavioral event, they often use motion picture film to record behavior (Eibl-Eibesfeldt, 1975). For example, ethologists studied the function of the "eyebrow flash" in social interactions by filming 67 hours of individuals in naturally occurring social situations (Grammer, Schiefenhoevel, Schleidt, Lorenz, & Eibl-Eibesfeldt, 1988). Across three different cultures, they found that the eyebrow flash, usually accompanied by a smile, was a universal social signal (e.g., a sign of saying "yes").

Once narrative records are created, researchers can study, classify, and organize the records. Particular hypotheses or expectations about the behaviors under observation can be tested by examining the data. 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 they are testing in the study.

Hartup (1974) obtained narrative records as part of his naturalistic study of children's aggression. He investigated different aspects of children's aggression, including the relationship between particular kinds of events that preceded aggressive behavior and the nature of the aggressive episodes that followed these precipitating events. 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 (1974) 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 inferences or impressions of the observers, individuals who were coding the narrative would not be influenced by what the observer inferred was going on. 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. Field notes are used by journalists, social workers, anthropologists, ethologists, and others. They do not always contain an exact record of everything that occurred. Events and behaviors that especially interest the observer are recorded and are likely to be interpreted in terms of the observer's specialized knowledge or expertise. For example, an ethologist might record how the behavior of one species appears to parallel that of another. Field notes tend to be highly personalized (Brandt, 1972), but they are probably used more frequently than any other kind of narrative record. Their usefulness as scientific records depends on the accuracy and precision of their content. Accuracy and precision 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, as well as 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. Adjang (1986) used a portable cassette recorder to make narrative records of his spoken observations of the teasing behavior of young chimpanzees; however, he then transcribed these spoken reports onto paper "as soon as possible" (p. 139). This rule is sometimes not easy to follow when observations are made in natural settings. An ethologist who is trying to record the behavior of animals in the wild is sometimes hampered by bad weather, animal migration, dwindling daylight, and so forth. Notes may have to be made quickly—even while the observer is quite literally on the run. At other times, observers may have to wait until they return to camp to make written records of behavior.

Decisions regarding what should be included in a narrative record must be made prior to observing behavior. We have seen, for example, that verbal narratives may differ in terms of the degree of observer inference that is appropriate or the completeness of the behavioral record. Thus, these aspects of a narrative record, as well as others, must be decided upon 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.

Quantitative Measures of Behavior

- Researchers often obtain quantitative measures such as frequency or duration of occurrence when they seek to describe specific behaviors or events.
- Quantitative measures of behavior use one of the 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).

Assume you want to do a naturalistic observation study investigating reactions to individuals with obvious physical disabilities by those who do not have such disabilities. In order to conduct your study, it would be necessary to define what constitutes a "reaction" to a physically disabled individual. Are you interested, for example, in helping behaviors, approach/avoidance behaviors, eye contact, length of conversation, or in some other behavioral reaction? As you consider what behaviors you will use to define people's "reactions," you will also have to decide how you will measure these behaviors. Assume, for instance, that you have decided to measure people's reactions by observing eye contact between individuals who do not have obvious physical disabilities and those who do. You would still need to decide exactly how you should measure eye contact. Should you simply measure whether a passerby does or does not make eye contact, or do you want to measure the duration of any eye contact? Of course, you may wish to use more than one measure, a research strategy that is recommended whenever it is feasible. The decisions you make will depend on the particular hypotheses or goals of your study. In making your decisions, you should take advantage of information you can gain by examining previous published studies that have used the same or similar behavioral measures. Thompson (1982), for example, has measured reactions to physically disabled individuals and found that the reactions frequently can be classified as unfavorable. We will now describe four general ways in which behavioral measures can be defined.

Key Concept

Measurement Scales Quantitative measures of behavior differ depending upon the scale of measurement you decide to use. Thus, it is important for you to be familiar with the types of measurement scales used in behavioral research. **Measurement scales** represent different levels at which behaviors can be quantified, and the different measurement scales influence how data are subsequently analyzed. There are four 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 4.1, and a detailed description of measurement scales is provided in Box 4.2. You will need to keep these four measurement scales in mind as you select statistical procedures for analyzing the results of the research you will be doing. 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. The observer could record on a checklist, for example, whether individuals make eye contact or do not make eye contact with a physically disabled person, whether children in a classroom are talking or are quiet, whether people use seat belts or do not use seat belts. Characteristics of participants—such as age, race, and sex—are also frequently recorded using a checklist, as are features of the setting—such as time of day, location, and whether other people are present. Researchers often are interested in observing behavior as a function of participant and context variables. Do males use seat belts more than females? Are people who drive inexpensive automobiles more likely to carpool than people who drive expensive automobiles? These questions can be answered by observing the presence or absence of certain behaviors (seat-belt use or carpooling) for different categories of participants and settings (male and female, expensive and inexpensive automobiles).

Tassinary and Hansen (1998) used an ordinal scale to measure male and female undergraduate students' reactions to line drawings of female figures. The figures varied on physical dimensions such as height, weight, and hip size. The undergraduates rank-ordered sets of figures in terms of attractiveness and fecundity (i.e., capability of bearing children). According to evolutionary psychology theory, female attractiveness is defined on the basis of cues simultaneously signaling both physical attractiveness and reproductive potential. One specific prediction based on evolutionary psychology theory is that the

TABLE 4.1 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 4.2

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.

waist-to-hip ratio should similarly predict *both* attractiveness and fecundity ratings. Contrary to this expectation, the results showed that relative hip size and weight were positively related to rankings of fecundity and negatively related to physical attractiveness ratings.

TABLE 4.2 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.

In order to quantify behavior in an observational study, observers sometimes make ratings of behaviors and events. Observers usually make ratings on the basis of their subjective judgments about the degree or quantity of some trait or condition (see Brandt, 1972). For example, Dickie (1987) asked observers to rate parent–infant interactions in the context of a study designed to assess the effects of a parent training program. Her observers visited homes and rated mothers and fathers while the parents interacted with their infant child. During most of the observation period, the observers sat in the room with the infant and asked the parents to "act as normal as possible—just as if we [the observers] weren't here." Parent–infant interactions were rated on 13 different dimensions, including degree of verbal, physical, and emotional interaction. For each dimension a continuum was defined that represented different degrees of this variable rated on a 7-point scale. A rating of 1 represented the absence or very little of the characteristic, and larger numbers represented increasingly more of the trait.

Table 4.2 outlines one of the dimensions used by the observers in this study: warmth and affection directed toward the child. Note that precise verbal descriptions are given with the four odd-numbered scale values to help observers define different degrees of this trait. The even-numbered values (2, 4, 6) are used by observers to rate events that they judge fall between the more clearly defined values. The investigators found that parents who had taken part in a program aimed at developing competency in dealing with an infant were rated higher than were untrained parents on many of the variables.

At first glance, a rating scale such as that used by Dickie would appear to represent an interval scale of measurement. There is no true zero, and the intervals seem to be equal. And, in fact, many researchers treat such rating systems as if they represent interval scales of measurement. Closer examination, however,

reveals that most of the rating scales used by observers to evaluate people or events on a psychological dimension really yield only ordinal information. For a rating system 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.

LaFrance and Mayo (1976) provide an example of ratio measurement in their study of racial differences in the amount of eye contact between individuals of the same race engaged in conversation. Pairs of Black individuals and pairs of White individuals were observed in natural settings. The amount of time each member of the pair spent looking into the face of the other member was recorded. Duration of eye contact represents a ratio level of measurement because units of time (e.g., seconds) have equal intervals and "zero" is a meaningul value (i.e., no eye contact). The researchers found that Blacks gazed less at another person while listening to conversation than did Whites. LaFrance and Mayo suggest that subtle differences in eye contact may be a source of social misunderstandings. White speakers may feel that lack of eye contact by a Black listener indicates untrustworthiness or lack of interest when it may reflect cultural differences between the races.

Another important measure of behavior is *frequency* of occurrence. Checklists can be used to measure the frequency of particular behaviors in the same individual or group of individuals by making repeated observations of the same individual or group over a period of time. The presence or absence of specific behaviors is noted at the time of each observation. 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 being counted, then zero represents an absence of that specific behavior. Ratios of scale values would be meaningful as long as, for instance, an individual with 20 units had twice as many units as someone with 10.

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 were outfitted with an ambulatory blood pressure monitor (Dolan, Sherwood, & Light, 1992). Students wore the electronic recording device during two "typical" school days, one of which, however, included an exam. Participants also completed a questionnaire assessing coping strategies and kept detailed logs of their daily activities. The researchers compared blood pressure readings for different times of the day and as a function of coping style. Students classified as exhibiting "high self-focused coping" (that is, who showed tendencies "to keep to themselves and/or blame

themselves in stressful situations," p. 233), had higher blood pressure responses during and after an exam than did those who were classified as low in self-focused coping strategies.

In the Experience Sampling Method (ESM), researchers supply participants with electronic pagers, usually for a week at a time. Participants are asked to report on their activities when signaled by the pagers (e.g., Csikszentmihalyi & Larson, 1987). In one study of adolescent interactions with their families, 220 middle- and working-class youth provided reports at two periods in their lives, grades 5 to 8 and grades 9 to 12 (Larson et al., 1996). When randomly "beeped," the adolescents used a checklist to indicate whom they were with, responded on 7-point scales regarding their emotional state and perceptions of the people they were with, and answered open-ended questions about their activities. As might be expected, the data revealed a sharp decline in the time adolescents spend with family members across grades 5 to 12. What was a surprising finding, however, was that amount of adolescent–family conflict was not related to changes in family time, something not seen in other primates where conflict frequently leads to disengagement from the family group. The researchers argued that for human adolescents, "the negative experience with family is as likely to be a stimulus for continued interaction—like one might see in an embattled, enmeshed family—as it is to be a stimulus for physical withdrawal" (p. 752).

Park, Armeli, and Tennen (2004) used an "Internet daily diary methodology" to obtain measures of stress and coping by University of Connecticut undergraduates. A total of 190 students participated by logging into a secure Internet website on each of 28 days. (E-mail reminders were sent each day.) Participants recorded information about the day's most stressful event. They then rated the event for how controllable they thought it was and answered questions about their mood and strategies of coping. Positive moods were linked more with problem-focused coping strategies than with avoidance strategies, especially when stressors were viewed as controllable. Other researchers have participants carry hand-held computers and make "electronic diary" notes when prompted (e.g., McCarthy, Piasecki, Fiore, & Baker, 2006; Shiffman & Paty, 2006).

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

Analysis of Observational Data

Data Reduction

- Observational data are summarized through the process of data reduction.
- Researchers quantify the data in narrative records by coding behaviors according to specified criteria, for example, by categorizing behaviors.

- Data are summarized using descriptive measures such as frequency counts, means, and standard deviations.

Analysis of Narrative Records Narrative records can provide a wealth of information about behavior in natural settings. Once data are collected, how do researchers summarize all this information? Data reduction is often an important step in analyzing the content of narrative records. **Data reduction** is the process of abstracting and summarizing behavioral data. Using *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. In qualitative analysis, data reduction occurs when researchers verbally summarize information, identify themes, categorize information, group various pieces of information, and record their own observations about the narrative records.

Key Concept

Data reduction often involves the process of **coding,** the identification of units of behavior or particular events according to specific criteria. For instance, Hartup's (1974) 10-week observation of children's aggression yielded information about 758 units of aggression. In a study of Internet postings related to adolescent self-injury behavior, researchers observed a total of 3,219 posts that were coded for various themes, such as motivation for self-injury and methods of concealing their behavior (Whitlock, Powers, & Eckenrode, 2006). As part of an ethological study of preschool children, McGrew (1972) identified 115 different behavior patterns. He developed coding schemes to classify 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. Coders used the coding schemes to classify these behavioral patterns while they watched videotape recordings of the observations that showed children attending nursery school.

Key Concept

Coding is often based on units of behavior or events that are related to the goals of the study. For instance, when coding observations of interactions between mothers and children from maltreating and nonmaltreating families, researchers identifed four types of maternal behavior and four types of child behavior (Valentino et al., 2006). Data reduction using coding allows researchers to determine relationships between specific types of behavior and the events that are the antecedents of these behaviors. For example, McGrew (1972) found that children exhibit a "pout face" after losing a fight over a toy. This ethologist observer noted that young chimpanzees show a similar expression when seeking reunion with their mother. Just after being frustrated (and often just prior to weeping), children exhibited a "pucker face." Interestingly, there seems to be no record of a pucker face in nonhuman primates.

Descriptive Measures Descriptive measures are used to summarize observational data when quantitative data analysis is used. When events are classified into mutually exclusive categories (nominal scale), the most common descriptive measure is relative frequency. A ratio of the frequency with which various behaviors occur over the total frequency of events observed is a relative frequency measure. Relative frequency measures are expressed as either a proportion or a percentage. For example, Jenni and Jenni (1976) observed students on

six college campuses. They reported that 82% of female college students carried their books by wrapping one or both arms around the books (with the short edges resting on their hip or in front of their body). Only 3% of male students used this particular book carrying method! Although book bags and backpacks are perhaps now the norm on college campuses, the next time you observe books actually being carried, take a look at the obvious differences between men and women in carrying behavior. Why do you think these differences exist?

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 a group. For a more complete description of group performance, researchers also report measures 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 12 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 4.3 gives the means and standard deviations summarizing the results of this study. The means in Table 4.3 show that White listeners spent more time looking into the faces of White speakers than Black listeners spent looking into the faces of Black speakers. 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

TABLE 4.3 MEANS AND STANDARD DEVIATIONS DESCRIBING THE TIME (IN SECONDS) THAT LISTENERS SPENT LOOKING INTO THE FACE OF A 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).

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.

Observer Reliability

- 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 discrepancies.
- 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.

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. One way researchers assess the reliability of an observer is to ask, "Would another observer viewing the same events obtain the same results?"

Key Concept

Interobserver Reliability The degree to which two independent observers agree is referred to as **interobserver reliability.** When observers disagree, we become uncertain about what is being measured and what behaviors and events actually occurred. Low interobserver reliability is likely to result when the event to be recorded is not clearly defined. Imagine Hartup (1974) asking his observers to record aggressive episodes among children without giving them an exact definition of aggression. What exactly is aggression? Some observers might decide to define aggression as one child's physical attack on another; other observers might include verbal assaults in their definition of aggression. What is a playful push and what is an angry shove? Similarly, precise definitions of "play" must be used by researchers observing this kind of child behavior (e.g., Valentino et a1., 2006). Without a clear definition of behavior or of the events to be recorded, observers do not always agree—and hence show low interobserver reliability. In addition to providing precise verbal definitions, giving concrete examples of a phenomenon generally helps increase reliability among observers. Showing photographs or videotapes of aggressive and nonaggressive episodes to observers would be a good way to improve their ability to classify aggressive behaviors reliably. Observer reliability is also generally increased by training observers and giving them practice doing the observations. It is especially helpful during the training and practice to give the observers specific feedback regarding any discrepancies between their observations and those of other observers (Judd, Smith, & Kidder, 1991).

A highly reliable observer does not necessarily make accurate observations. Consider two observers who reliably agree about what they saw but

who are both "in error" to the same degree. Neither observer is providing an accurate record of behavior. For example, both might be influenced in a similar way by what they expect the outcome of the observational study to be. Instances are occasionally reported in the media of several observers claiming to see the same thing (for instance, an unidentified flying object, or UFO), only to have the event or object turn out to be something other than what observers claimed it to be (for instance, 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 outcome expectancies, fatigue, or boredom is generally so small that we can be confident that what was reported actually occurred. Of course, the more independent observers agree, the more confident we become.

Measures of Reliability 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

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

Hartup (1974) reported measures of reliability using percentage agreement that ranged from 83% to 94% for judges who coded narrative records according to type of aggression and nature of antecedent events. When observing the simple act of smoking among adults and youth, observers averaged 99% (Jason et. al., 2006). Although there is no hard-and-fast percentage of agreement that defines low interobserver reliability, researchers generally report estimates of reliability that exceed 85% in the published literature, suggesting that 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 use only a sample of the observations to measure reliability. For example, two observers might be asked to record behavior according to a time-sampling procedure such that there is only a subset of times during which both observers are present. Amount of agreement for the times when both observers were present can be used to indicate the degree of reliability for the study as a whole.

When observational data represent at least an interval or ratio scale, such as when time is the variable being measured, 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 of the 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 engaged 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. As noted on previous page, when interval or ratio data are collected, a Pearson correlation coefficient, *r*, may be used to obtain a measure of interobserver reliability. This measure tells us how well ratings of two observers agree.

The correlation coefficient indicates the *direction* and *strength* of the correlation. 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. (Measures of smoking and lung cancer are positively correlated.) A negative correlation indicates that as the value of one measure increases, the value of the other measure decreases. (Time spent watching television and scores on academic tests are negatively correlated.) Clearly, when assessing interobserver reliability, we are looking for positive correlations. The strength of a correlation refers to the degree of covariation present (or, as is sometimes said, the strength of the predictive relationship, since correlation is the basis for making predictions about behavior and events). (This aspect of correlation is discussed more fully in Chapters 5 and 12.) Correlation coefficients 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 (and hence no basis for making predictions). 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 the 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 greater the better!). As you have seen, LaFrance and Mayo (1976) reported average interobserver correlations of .92, showing very good agreement between their observers.

In Chapter 12 we discuss correlations more fully, 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 the material on correlation in Chapter 12.

Thinking Critically About Observational Research

Influence of the Observer

- If individuals change their behavior when they know they are being observed ("reactivity"), their behavior may no longer be representative of their normal behavior.
- Research participants may respond to demand characteristics in the research situation to guide their behavior.
- Methods to control reactivity include unobtrusive (nonreactive) measurement, adaptation (habituation, desensitization), and indirect observations of behavior.

- Researchers must consider ethical issues when attempting to control reactivity.

Conducting a good observational study involves choosing how to sample behavior and events to observe, choosing the appropriate observational method, and choosing 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 occurs because people often change their behavior when they know they are being observed. A second problem occurs when observers' biases influence what behavior they chose to record. We'll consider each of these problems in turn.

Key Concept

Reactivity The presence of an observer can lead people to change their behavior because they know they are being observed. When the observer influences the behavior being observed, the problem of **reactivity** is present. When individuals "react" to the presence of an observer, their behavior may not be representative of 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 located himself on the street corner with clipboard in hand. He soon noticed that all the cars were stopping at the stop sign. He then realized that his presence was influencing the drivers' 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 a little anxious about participating in psychological research. Measures of arousal, such as heart rate and galvanic skin response (GSR), may show changes simply as a function of an observer's presence. Wearing an electronic beeper that signals when to record behavioral activities and mood also can be expected to affect participants' behavior (e.g., Larson, 1989).

Key Concept

Research participants 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 cues present in the setting, the research procedure itself, and implicit cues given by the researcher. As participants try to guess what is expected, they may change their behaviors accordingly. Participants' responses to the demand characteristics of a research situation pose a threat to the external validity of psychological research. Our 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 psychological research setting. Interpretation of the study's findings is potentially threatened because participants may unintentionally make a research variable

look more effective than it actually is or even nullify the effects of an otherwise significant variable. The problem of demand characteristics can be reduced by limiting individuals' knowledge about their role in a study or about the hypothesis being tested in the study. The researcher's goal in keeping participants unaware of important details regarding the study is to obtain more representative behavior. This methodological "solution" to the problem of demand characteristics does raise ethical concerns about important issues such as informed consent.

Controlling Reactivity There are several approaches that researchers use to control the problem of reactivity. They can eliminate reactivity by making sure that research participants do not detect the presence of the observer. Measures of behavior when participants do not know they are being observed are referred to as *unobtrusive (nonreactive) measures,* most often used in naturalistic observation. Obtaining unobtrusive measures may involve concealing the observer or hiding mechanical recording devices such as tape recorders and videotape cameras. LaFrance and Mayo (1976) observed people in conversation without their knowledge. Observations were made in a variety of natural settings, such as restaurants and waiting rooms, so we can imagine that observers had to keep their stopwatches and data sheets hidden behind menus and potted plants in order to obtain unobtrusive measures of behavior. Yet another approach is for researchers to use disguised participant observation and to adopt a role in the situation other than that of observer. You may remember this procedure was used by social psychologists studying the behavior of individuals who claimed to be in contact with aliens from outer space (Festinger et al. 1956). When researchers use unobtrusive measures, they assume that participants act as they ordinarily would because the participants do not know that an observer is present.

Another approach researchers use to deal with reactivity is to adapt participants to the presence of an observer. Researchers make a reasonable assumption that as participants get used to an observer being present, they will come to behave normally in that person's presence. Adaptation can be accomplished through either habituation or desensitization. In a *habituation* procedure, observers simply introduce themselves into a situation on many different occasions until the participants cease to react to their presence. In order to film a documentary titled *An American Family,* which was shown on public television in the early 1970s, observers (with their cameras) literally moved into a California home and recorded the activities of a family over a 7-month period. Although it is impossible to tell how much of their behavior was influenced by the observers' presence, the events that unfolded and remarks made by family members provided evidence that a habituation process took place. During filming, the family broke up, the mother asking the father to move out of the house. When interviewed later about having the divorce announced to millions of television viewers, the father admitted that they could have asked the camera crew to get out but that, by this time, "we had gotten used to it" (*Newsweek,* 1973, p. 49).

Desensitization as a means of dealing with reactivity is similar to the desensitization used in the behavioral treatment of phobias. In a therapy situation, an

FIGURE 4.4 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.

individual with a specific fear (say, an irrational fear of spiders) is first exposed to the feared stimulus at a very low intensity. The patient may be asked to think of things that are related to spiders, such as dusty rooms and cobwebs. At the same time, the therapist helps the patient practice relaxation. Gradually the intensity of the stimulus is increased until the patient can tolerate the actual stimulus itself. Desensitization is often used by ethologists to adapt animal subjects to the presence of an observer. Prior to her violent death in the land of her beloved subjects, Fossey (1981, 1983) conducted fascinating observational studies of the mountain gorilla in Africa. Over a period of time she moved closer and closer to the gorillas so that they would get used 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 the research equipment she was using.

Finally, nonreactive measures of behavior can be obtained by observing behavior indirectly (see Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). This may involve examining physical traces left behind or examining archival information, which are records kept by society about individuals and events. 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 4.4). Another researcher used the records kept by a library to assess the effect on a community of the introduction of television. Withdrawals of fiction

titles dropped, but the demand for nonfiction was not affected (see Webb et al., 1981). Physical traces and archival data are important unobtrusive measures that can be valuable sources of information about behavior. These methods will be introduced in greater detail in Chapter 6.

Ethical Issues Whenever researchers try to control for 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 (as we discussed in Chapter 3) and must include a consideration of the sensitivity of the information, the setting where observation takes place, and the method of dissemination of the information obtained (e.g., Deiner & Crandall, 1978).

Recent behavioral studies using the Internet introduce new ethical dilemmas. When researchers entered Internet chat rooms as disguised participant observers to find out what makes racists advocate racial violence (Glaser, Dixit, & Green, 2002), the information they obtained could be seen as gathering incriminating evidence without the respondents' knowledge, not unlike 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 the chat room constituted a "public forum," that these topics were common to that forum, and that the researchers had instituted sufficient safeguards to protect the respondents' identities (e.g., by carefully separating names, typically the pseudonyms commonly used by individuals in this chat room, from the responses). On the other hand, there are instances in which people have felt that 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 posts about self-injurious behaviors were required by their university IRB to use paraphrases of participant quotes rather than exact quotes (Whitlock, Powers, & Eckenrode, 2006). Behavioral research using the Internet is just beginning, and both researchers and IRB members are in a learning phase. Creative problem solving will be required by both groups if these ethical dilemmas are to be resolved (see Kraut et al., 2004).

When individuals are involved in situations that are deliberately arranged by an investigator, as might happen in a structured observation or in a field experiment, ethical problems associated with placing participants at risk may arise. Consider, for instance, a study designed to investigate how college students' attitudes toward racial harassment are affected by hearing other students either condone or condemn racism (Blanchard, Crandall, Brigham, & Vaughn, 1994). More than 200 White undergraduate women attending various universities were "naive participants." The women were approached by a White interviewer as they walked across campus and were invited to answer a short series of questions about "how their college should respond to acts of racism" (p. 994). A female confederate, posing as a student, approached the interviewer so that she arrived at the same time as the naive participant. The interviewers asked both "students" the same five questions; however, the interviewer always

questioned the confederate student first. At this point, the confederate responded by either condemning or condoning racists' acts. Of interest was the effect of these statements on the naive participants' responses to the same questions. The results were clear: Hearing another student condemn racism produced more condemning responses relative to a no-influence control group, and hearing another student condone racism produced more condoning reactions to racism than hearing no one else express an opinion. Thus, as the authors suggest, the findings "imply that a few outspoken people can influence the normative climate of interracial social settings in either direction" (p. 997).

Were the naive participants "at risk"? If you think the participants were at risk, what degree of risk was present? Did the goals of the study, and the knowledge potentially obtained, outweigh the risks involved in the study? Although participants were "debriefed immediately" in this study, is that sufficient to address any concerns that the naive students might have about how they behaved when confronted with racist opinions, or even to restore confidence in a science that seeks knowledge through deception? Attempting to provide answers to these kinds of questions highlights the difficulty of ethical decision making. (In responding to these ethical questions, it may be helpful to refer to the recommended steps in the process of ethical decision making that are outlined at the end of Chapter 3.)

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 the 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.

Earlier in this chapter we described a study in which Rosenhan (1973) and his colleagues observed the interaction between staff members and patients in mental hospitals, and they found a serious bias on the part of the staff. Once patients were labeled schizophrenic, their behavior was interpreted in light of this label. Staff members interpreted behaviors that might have been considered normal when performed by "sane" individuals as evidence of the patients' insanity. For instance, the researchers later learned that note-taking by the participant observers, which was done openly, had been cited by members of the staff as an example of the pseudopatients' pathological state. Thus, the staff tended to interpret patients' behavior in terms of the label that had been given them. This example clearly illustrates the potential danger of **observer bias,** the systematic errors in observation that result from an observer's expectations.

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. This expectancy may be created by knowledge of the results of past investigations or perhaps by the observer's own hypothesis about behavior in this situation. Expectancies can be a source of observer bias—*expectancy effects*—if they lead to systematic errors in observation (Rosenthal, 1966, 1976). Cordaro and Ison (1963) designed a study to document expectancy effects. The study required college student observers to record the number of head turns and body contractions made by two groups of flatworms. The observers were led to expect different rates of turning and contracting in the two groups. The worms in the groups were, however, essentially identical. What differed were the observers' expectations about what they would see. Results showed that the observers reported twice as many head turns and three times as many body contractions when a high rate of movement was expected than when a low rate was expected. Apparently, the students 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 movie cameras would eliminate observer bias. Although automation reduces the opportunity for observer bias, it does not necessarily eliminate it. Consider the fact that, in order to record behavior on film, the observer must determine the angle, location, and time of filming. To the extent that these aspects of the study are influenced by personal biases of the observer, such decisions can introduce systematic errors into the results. Altmann (1974) describes an observational study of animal behavior in which the observers biased the results by taking a midday break whenever the animals were inactive. Observations of the animals during this period of inactivity were conspicuously absent from the observational records. Furthermore, using 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.

Observer bias also can be reduced by limiting the information provided to observers. When Hartup (1974) analyzed the results of his observational study of children's aggression, the individuals who performed the analysis were not permitted to see all the narrative records. When the nature of the aggressive act was classified, the antecedent events were blacked out; and when antecedent events were coded, the nature of the aggressive act was blacked out. Therefore, in making their classifications, the coders could not be influenced by information related to the event that they were coding. In a manner of speaking, the coders were "blind" to certain aspects of the study. Observers are *blind* when they do not know why the observations are being made or the goals of a study. When trained coders analyzed the videotapes of interactions between mothers

and children from maltreating and nonmaltreating families, they were not aware of what type of family they were observing (Valentino et al., 2006). Using blind observers greatly reduces the possibility of introducing systematic errors due to observer expectancies.

SUMMARY

Researchers can 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. Observational methods can be classified on two dimensions: the degree of observer intervention and the manner in which behavior is recorded. Observation in a natural setting without observer intervention is called naturalistic observation. Observation with intervention can take the form of participant observation, structured observation (frequently used by developmental psychologists), and field experiments (often used by social psychologists). In an observational study, 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 are typically used when researchers are interested in whether a specific behavior has occurred (and under what conditions). Frequency, duration, and ratings of behaviors are common dependent variables in observational studies.

How quantitative data are described and analyzed depends on the scale of measurement used. The four measurement scales used by psychologists are nominal, ordinal, interval, and ratio. When narrative records are made, some type of coding system is generally used as one step in the process of data reduction. Measures of frequency and duration, as well as ratings, are typically summarized using descriptive statistics such as 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 that has been 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. Finally, researchers must address ethical issues prior to beginning a research study. Ethical issues are especially salient when an observational study involves a form of deception such as disguised participant observation or the use of unidentified confederates. Internet research raises new ethical dilemmas that need to be addressed by both researchers and IRB members (e.g., privacy issues in chat rooms).

KEY CONCEPTS

external validity 97
time sampling 97
situation sampling 98
naturalistic observation 100
participant observation 103
structured observation 106

field experiment 109
narrative records 111
measurement scale 114
data reduction 119
coding 119
interobserver reliability 121
correlation coefficient 123
reactivity 124
demand characteristics 124
observer bias 128

Review Questions

1 Identify three characteristics of scientific observation that distinguish it from our everyday observation.
2 Explain why researchers use sampling in observational studies, and describe what the proper use of sampling is intended to accomplish.
3 Explain how the degree of intervention and the method of recording behavior can be used to classify observational methods.
4 Describe a research situation in which naturalistic observation can be useful when ethical considerations prevent researchers from controlling aspects of human behavior.
5 Identify three factors in participant observation that researchers need to consider to determine the extent of the observer's influence on the behavior being observed.
6 Structured observation represents a compromise between naturalistic observation and laboratory experiments. What are the primary advantage and potential cost of this compromise?
7 Give an example using each of the four measurement scales of how a researcher could measure eye contact between pairs of people in conversation with each other.
8 What are the most common descriptive measures (a) when events are measured on a nominal scale and (b) when behavior is recorded on at least an interval scale?
9 Describe the effects of each of the following factors on interobserver reliability: definition of the event being observed, training, practice with feedback.
10 What two types of information do you gain by knowing the sign and the numerical value of a correlation coefficient?
11 Identify the measurement scales that require a correlation coefficient to assess interobserver reliability, and explain what a negative correlation would indicate in this situation.
12 Explain whether high interobserver reliability ensures that the observations are accurate and valid.
13 Explain why participants' reactions to demand characteristics can be a threat to the external validity of psychological research and to the interpretation of a study's findings.
14 Explain how the ethical issues of privacy and risk can arise when researchers use unobtrusive measures such as concealing the presence of an observer.
15 What new ethical considerations are involved when research is conducted on the Internet?
16 Describe two ways in which observer bias (expectancy effects) can occur in psychological research.
17 What is the best procedure to 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 do 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 friend of yours is absolutely convinced that he has a positive influence on the friendliness of conversations in which he is a participant. He has reached this conclusion on the basis of his everyday observations. You convince him that a systematic study is needed to confirm his hypothesis. Your friend (still smiling) carefully develops an operational definition of the friendliness of a conversation and records a rating for each of the next 50 conversations in which he is a participant. His results show that 75% of these conversations are rated "very friendly," 20% are rated "friendly," and 5% are rated "neutral." Your friend returns to you—now convinced beyond a shadow of a doubt that he has a positive effect on the friendliness of a conversation. Although your friend won't be pleased with you, explain to him why his study is seriously flawed as a basis for confirming his hypothesis. In your critique of your friend's study, be sure to address the following issues related to the study:

- **A** Explain what comparison set of observations he needs before he can begin to make a claim that he has a positive influence on the friendliness of conversations. Explain how this comparison set

of observations could even lead to the conclusion that your friend has a negative effect on the friendliness of conversations.

B Explain why your friend should not be the one to select the conversations to observe nor the one who records how friendly the conversations were.

C Does your critique of your friend's study allow you to conclude that your friend does not have a positive effect on conversations in which he is a participant? Why or why not?

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 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 The students used naturalistic observation in this study. They did not intervene in the situations they were observing, and nonintervention is a defining characteristic of naturalistic observation. The study also was done in natural settings, the library and the student union.

2 The students' choice to use a 5-minute observation period may have limited their ability to measure students' concentration effectively. The 5-minute observation period (especially if the observers began timing the 5-minute interval when the student was looking at the material or writing) may have been too short to "show" changes in concentration. If the 5-minute interval was too short to show changes in concentration, then it is unlikely that differences in concentration would have been observed between the two locations. One possible way to improve the operational definition would be to lengthen the observation interval to 20 or 30 minutes.

3 Students' ability to concentrate may vary across days of the week and times of the day. The students chose to observe on a Monday evening from 9 to 11 P.M. because they thought they would be observing at a "prime" study time for students. The time-sampling plan could be improved to increase external validity by making observations at different times of the day, on different days of the week, and across the weeks of the semester.

4 If we assume that students can concentrate better in the library than in the lounge in the student union, then we need to find some way to explain why study times were the same in this study across the two locations. One possibility is that students chose to study different material in the two locations. If students in the student union were studying "easier" material, then they could have concentrated as well as students who were studying harder material in the library. One of the challenges in doing naturalistic observation is that the researcher cannot control factors that possibly could influence the outcome of the observations (because of the nonintervention that characterizes naturalistic observation).

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 suggested that they were sensitive to the problem of reactivity. Two other methods they might have used are habituation and desensitization.