# Chapter 6

## **Methods of Data Collection**

### **Introduction to Methods of Data Collection**

#### The Nature of Observations

## **Ways of Observing**

Participant vs. Nonparticipant Observation

**Scheduling Observations** 

Defining the Behavior to Be Observed

Specific Techniques for Recording Behaviors

Recording More Than One Response

Thinking Critically About Everyday Information

## **Reliability of Observations**

Interobserver Agreement

Measuring the Reliability of Observational Data

# **Recordings by Equipment**

**Public Records** 

# **Survey Methods**

Questionnaires

**Instruments and Inventories** 

Interviews

Laboratory vs. Field Research

**Case Analysis** 

**General Summary** 

**Detailed Summary** 

**Key Terms** 

**Review Questions/Exercises** 

### **Introduction to Methods of Data Collection**

By now, it should be abundantly clear that behavioral research involves the collection of data and that there are a variety of ways to do so. For example, if we wanted to measure aggressive behavior in children, we could collect those data by observing children with our eyes, by using equipment to measure the force with which they hit an object, by examining juvenile crime records, by surveying parents and teachers, by interviewing parents and teachers, or by administering an aggression scale to children. This is just a sample of the methods that are possible; we are sure that you could imagine many others. However, these examples do illustrate several distinctly different methods that can be used to collect data. As with most research design techniques, each method has advantages and limitations. Perhaps the most interesting and challenging of these is the method of observation. (In a sense, all of behavioral research is based upon observation. What we describe here is a specific kind of observational procedure.) Historically, behavioral research has relied heavily on this method, and it will undoubtedly continue to be a primary method for gathering behavioral data. Let's begin this chapter by discussing the nature of observations, the ways to make observations, and the reliability of observations. Following this discussion, we will turn to other methods illustrated by the examples above.

#### The Nature of Observation

Observations involve the use of our sensory systems (including eyes and ears) to record behavior. They require that humans make *judgments* about the occurrence of the behavior, its frequency, its duration, or its latency. These measures are the basic data used to describe naturally occurring behavior or to assess the effects of our independent variable. Therefore, they must be gathered with care. Observations such as these (judgments that behavior has or has not occurred) are inherently more subjective than other data collection procedures. Because judgments are based on our perceptions, the same event occurring in the environment will be perceived differently by different people. We may observe Johnny chasing Mary around the playground and call it aggressive behavior. You may observe the same event and call it play behavior. Certainly, attention is one important factor. I may be attending to the conversation between two people, and you may be attending to the body language exhibited by those two people. As a result, we may interpret our observations similarly or differently.

Behavior measures based upon response categories such as play, aggression, and self-assertion are more complex than specific responses such as talking, walking, or attending. These, in turn, are very

different from the lever press, the key peck, heart rate, latency to respond, or eye blink responses. The recording of these behaviors is usually automated, and there is little doubt about the criteria used for their occurrence or nonoccurrence. They do not require the judgment of human observers.

Data derived from human observers are playing an increasingly important role in research, particularly in applied settings. They also play an important role in some laboratory settings. As you know, the interests of psychologists are extremely varied. This broad scope of interest has resulted in studying behavior under a wider variety of conditions than in the past. Some studies involve making unobtrusive observations of animals behaving in their natural settings, including primates and animals lower on the phylogenetic scale. Psychologists often record laboratory observations in addition to the behavior that is automatically recorded. These observations are often interesting behaviors in their own right. Applied psychologists sometimes encounter special problems. For example, behavior modification experts must be adept at recognizing when a given behavior achieves criterion (for example, when the sound made by a mute autistic child is sufficiently close to "mama" to warrant rewarding the child). Moreover, observers must be able to make fine distinctions between behaviors that are closer and more distant approximations to the criterion behavior. Similarly, those studying the effects of psychologically active drugs must attend to a wide variety of both specific responses and broader classes of behavior. Reliable observations are essential in assessing the treatment effects. In many instances, the primary and often the only data are derived from direct observation by the investigators. It is essential that these primary data be accurate, objective, and reliable.

In this chapter, we will discuss some problems associated with using human observers and some factors that decrease the accuracy and reliability of observations. We will also discuss some ways of assessing interobserver reliability. You should be aware that the method of observation does not apply to only some research designs. In fact, all of the research designs discussed in this book, both experimental and nonexperimental, can involve observation. One specific type of nonexperimental design, naturalistic observation, always involves observation; it will be discussed in Chapter 15.

# Ways of Observing

#### Participant vs. Nonparticipant Observation

There are two broad approaches to observing behavior: **participant observation** and **nonparticipant observation**. Participant observers conduct their observations "from the inside"; that is, the researcher is an integral part of the environment being observed. Nonparticipant observers conduct their observations "from the outside"; the researcher does not interact with those being observed.

The following extract describes the habituating techniques used by Jane Goodall to study the mountain gorilla in its natural habitat:

My technique of habituating the gorillas was simple but essential, for I could only obtain unbiased data on their behavior if they remained relatively unaffected by my presence. I usually attempted to approach the group undetected to within about 150 feet before climbing slowly and in full view of the animals onto a stump or the low branch of a tree where I settled myself as comfortably as possible without paying obvious attention to them. By choosing a prominent observation post not only was I able to see the gorillas over the screen of herbs, but, reciprocally, they could inspect me clearly, which was the most important single factor in habituating the animals. Under such circumstances they usually remained in the vicinity to watch me, and even approached me to within 5 feet. (Schaller, 1963, p. 22)

This description of Jane Goodall's methodology suggests that she was a nonparticipant observer. However, if you are familiar with her work with gorillas, you know that the gorillas not only habituated to her presence but later began to interact with her. At this point, she became a participant observer.

Let us imagine that we want to observe 9-year-old children on the playground and that we want to record frequency of aggressive behavior. After establishing an operational definition of aggressive behavior, we must decide how the observations will be made. Several options are available. As a participant observer, you could play with the children on the playground and make them aware that you are observing their behavior. Another option as a participant observer is to play with the children on the playground without making them aware that you are observing their behavior. A third option is to be a nonparticipant observer, observing the children "from a distance" and making them aware that you are observing their behavior. A final option is to be a nonparticipant observer, observing the children "from a distance" without their awareness. Note that "from a distance" may refer literally to a geographic distance or may represent a method such as a one-way mirror or a hidden camera. So, which method would you use? If you are already considering the pros and cons of each approach, then you are thinking like a behavioral scientist!

Whether you are observing gorillas in the rainforest or children on the playground, you will probably not be able to observe them continuously. Scheduling observations is therefore an important consideration.

#### **Scheduling Observations**

Decisions must be made regarding frequency, duration, and time of day for scheduling observations. These decisions depend upon the purposes to be served by the observations. Researchers may only be interested in a restricted, narrow time period, or they may be interested in a broad, representative time

period. If you are interested in a representative description of naturally occurring behavior, it is necessary to observe over many different time periods and across several days. Jane Goodall's description of primate behavior under natural conditions is an excellent example. Laboratory studies, in contrast, often follow a rigid schedule in that the experiment takes place at the same time each day for an hour or more. In this case, the observation period is confined to the laboratory session time. An applied program would follow a different approach. A behavior modification program dealing with disruptive and unruly behavior in social settings would require that observations be made in each setting where the disruptions occur. Whatever the purpose of the observations, individual observation sessions should be sufficiently long to provide an adequate sample of the response of interest. Behaviors that occur with a low frequency may require longer observation periods than high-frequency behaviors. For some purposes, it may be necessary to sample a behavior at different times during the day to determine its range and variability. For most studies, however, observations are made at the same time each day so that observations are made under similar conditions from day to day.

If the occurrence or nonoccurrence of complex behavior is being judged, the criteria for establishing the presence of this behavior must be established. Before the behavior is recorded as present, these criteria must be satisfied. To minimize the observer's drifting away from the originally established criteria, a periodic review should be scheduled during the course of observing. Although **observer drift** has been a problem in some studies, it is usually correctable.

#### Defining the Behavior to Be Observed

As we have noted, our concern is to make observations that are both objective and reliable. We want to maximize "pure" observations and minimize the degree to which our observations are affected by our own interpretations and inferences. To this end, we attempt to define the dependent variable (behavior) in terms of specific observable responses and to specify clearly the criteria for judging when the behavior has occurred. As discussed in Chapter 5, we need clear operational definitions for behaviors to be observed. This is an important step if we are to ensure that different observers make similar observations. A good response measure will have relatively high interobserver agreement. The more precise we are in specifying our definition of a response or of the behavioral criteria, the higher the interobserver agreement will be. For example, if we were interested in the self-mutilating behavior of an autistic child, it would be unsatisfactory to instruct the observers to record every instance of self-mutilating behavior. The term is too broad, abstract, and undefined to be useful or to assure agreement among observers. However, if we operationally defined self-mutilating behavior as "head banging," "biting one's body," or "beating oneself with fists," we could more easily and reliably measure this class of behavior. Similarly, if observers are instructed to record incidents of children's aggressive behavior on the playground, one observer's

perception of aggression could be quite different from another observer's perception of aggression. Again, an operational definition is needed so that we know specifically what is being measured.

#### Specific Techniques for Recording Behaviors

Consider the following. A research team is interested in observing the self-mutilating behavior of children diagnosed as autistic. Having decided on a satisfactory operational definition of the target behavior, the team must now decide on the observational technique to use. Three choices are commonly available: (1) count the number of occurrences of self-mutilating behavior during an observational session (frequency method); (2) record the period of time during which the target behavior lasts (duration method); or (3) break the observational sessions into equal time intervals and record the occurrence of self-mutilating behavior within each interval (interval method). Let's take a closer look at each technique.

Frequency Method. The frequency method of observation is simple, straightforward, and easily understood. The observer simply counts the number of occurrences of the behavior of interest in a given interval of time. The interval of time is arbitrary; it may be as little as a few minutes or as much as several hours. Further, it may be based on one observation session, or it may run across many sessions over several days. Often, the investigator is only interested in the frequency of the observations in a given session and not in changes that may occur during that session. However, when the interest centers on assessing change or obtaining interobserver agreement within a session, the session may be divided into smaller, equal time intervals. For example, a 60-minute session may be divided into ten 6-minute intervals, and changes in frequency can be observed across this entire interval. Usually, observation periods are of the same duration from day to day. If not, then comparing frequencies based on 15 minutes with those based on 30 minutes would not be appropriate. If different durations are used, it is necessary to convert frequency to response rate. This measure can easily be derived by dividing the frequency of the response by the unit of time—for example, responses per minute. The frequency method of recording observational data is most appropriate with discrete responses that take a relatively constant period of time to complete—for example, number of cigarettes smoked, number of words spoken, or number of head-banging incidents.

Responses occurring over long periods of time would not be appropriate for the frequency method. Responses such as time spent talking, sleeping, eating, or observing would be inappropriate. Counting these observations would be wasteful of information, and counting their frequencies might not be a sensitive way to assess a treatment effect. For these responses, the duration method is more appropriate.

Frequency measures, particularly rate measures, are popular among psychologists interested in the experimental analysis of behavior. In applied settings, attempts are often made to modify both excesses

(such as fighting) and deficits of behavior (such as not talking). In these settings, monitoring frequency is obviously important. Moreover, frequency and rate measures have been shown to be sensitive to contingencies of reinforcement.

**Duration Method.** As we noted, in those instances where response duration is long and/or the occasions of its occurrence relatively infrequent (such as sleep or inactivity), it is usually inappropriate to record frequency. Instead, we want to measure the duration of the response. For example, using the **duration method of observation,** we might measure the duration of time spent either in solitary or in social activity. Using a stopwatch or an event recorder, the observer activates the instrument when the behavior begins and terminates the time recording when the behavior ends. Even though response frequency is not of primary interest, it is possible to obtain a frequency measure by simply counting the number of times the recording instrument was activated. Judging when a response is initiated or terminated can often be difficult. For example, imagine you are studying the talking of an autistic child. There may be many pauses of varying duration, brief interruptions, or changes in the intensity of the behavior that requires a judgment as to whether a different response occurred. It is important that another observer (interobserver agreement) be able to make similar observations. Consequently, the decisions must be made as objective as possible.

Interval Method. The interval method of observation is the most flexible and widely used recording method. This method permits the recording of any behavior, whether discrete (head banging) or continuous (sleeping). With this method, the observation period is broken into equal intervals, the size of which varies with the particular observations of interest. Behavior is recorded as occurring (yes) or not occurring (no) in each interval. The interval size may be as short as a few seconds or as long as a few minutes, depending on the behavior under observation. A desirable time interval is long enough to accommodate a single response but not long enough to accommodate more than one response. Research interest is focused on the number of intervals during which the response occurs. For high-rate, short-duration responses, the interval should be short so that no more than one response per interval will occur. If more than a single discrete response can occur per interval, then counting the intervals may underestimate the frequency of the observed behavior. In contrast, if the intervals are so short that a single response can fall into two or more intervals, counting the intervals in which a response occurs may overestimate the frequency of behavior.

Obviously, the criterion for occurrence or nonoccurrence of behavior must be clear so that its occurrence can be quickly and reliably determined. The problem is more difficult than it at first appears because only a portion of the behavior may occur in a given interval. It becomes even more difficult if the

observer is recording several different responses concurrently. When this happens, a decision rule is sometimes adopted; if a response fills one-half or more of the interval, it is scored as occurring in that interval. For continuous responses (such as talking), each interval in which talking occurred would be scored as an occurrence. Therefore, the interval scoring method allows the investigator to derive both frequency (discrete responses) and duration (continuous responses) data.

#### Recording More Than One Response

With the interval method, it is common to score several different responses that occur concurrently in an interval. When this is done, it is essential to reserve time at the end of each interval for recording whether the different behaviors occurred. We will first give an example of scoring one response across different participants. Then, we will give an example of scoring multiple responses.

Let us assume that our observational period will be daily 30-minute sessions and we are interested in the occurrence or nonoccurrence of a single behavior. We decide that an interval size of 30 seconds is appropriate for our response. Therefore, we would divide the 30-minute session into 60 30-second intervals. For each of these 30-second intervals, we would record whether or not the behavior occurred. If we were observing three participants, our scoring sheet would look like Table 6.1. Please note the documentation required for each scoring sheet. This is important information that can easily be forgotten if not recorded in permanent form.

Table 6.1	Inte	erval Scoring Sheet for a Single Response With Three Participants									
Experiment:				Researcher:			Ol	Observer:			
Participants:				Location:			Da	Date:			
Time:				Behavior:			Co	Codes:			
			•				•				
				Intervals							
		1	2	3	4	5	6	7		60	
Participant 1		+ -	+ -	+ -	+ -	+ -	+ -	+ -		+ -	
Participant 2	2	+ -	+ -	+ -	+ -	+ -	+ -	+ -		+ -	
Participant 3	3	+ -	+ -	+ -	+ -	+ -	+ -	+ -		+ -	

Because we are recording only one response or one kind of behavior whenever it occurs during the interval, we can circle the + as soon as it is observed. If, by chance, the response occurs again in that same interval, we would simply ignore it. Finally, we would circle the minus sign if the behavior did not occur during the interval.

When several participants are observed during each session, only a single participant is observed at any given time. Thus, if there were 60 intervals, each participant would be observed independently 60 times. You could start with Interval 1, Participant 1, go to Interval 1, Participant 2, then to Interval 1, Participant 3. Then you could begin the sequence again with Interval 2.

When more than one behavior is being observed and recorded, it may be necessary to reserve a part of the interval for scoring purposes. Otherwise, the time it takes to record the occurrence or nonoccurrence of several responses may interfere with the task of observing. As a result, some responses may be missed. It is not difficult to reserve time for recording purposes. With our 30-second interval, we could designate the first 25 seconds for observing and the last 5 seconds for recording. If we were observing smiling (S), frowning (F), hitting (H), and biting (B), our data sheet might look like Table 6.2. If any of the behaviors occurred during the 25-second observation period, we would record it during the 5-second recording period simply by circling the proper code.

Table 6.2         Interval Scoring Sheet for a Single Participant and Four Different Responses							
Experiment:		Researcher:		Observer:			
Participant:		Location:		Date:			
Time:		Behavior: smiling, fr hitting, biting	owning,	Codes: S, F, H, B			
			Intervals				
	1	2	3		60		
Participant	SFHB	S F H B	SFHB		S F H B		

We should note that when short intervals are used, it may be necessary to use a signaling device to identify the beginning and end of each interval. A cassette recorder with an earpiece can fill the bill admirably. The precise time intervals can be recorded prior to making observations, and listening to the tape can pace the observers through the session, telling them exactly when to observe and when to record. With time intervals of 30 seconds, the tape could be arranged to signal the beginning of the observation interval by prerecording "Time Interval 1, Begin" and 25 seconds later "Time Interval 1, Record." Then the tape would identify "Time Interval 2, Begin," and so on. This procedure could be used when there is more than one observer. By coupling the recorder with two or more earpieces, we would have an excellent means of assuring ourselves that the investigators are observing and recording the same time periods. This achievement is particularly important to assess interobserver agreement.

After the experiment is completed, the interval observations can be converted into percentages. This is done by taking the number of intervals in which the response occurred, dividing it by the total number

of intervals, and then multiplying by 100. Thus, if a response was observed during 12 out of 60 observational periods, the percentage would be:  $12 / 60 \times 100 = 20\%$ .

Before we consider the issue of reliability of observations, let's think about an MSNBC news report of a study that examined poor posture associated with the use of school backpacks by children (see "Thinking Critically About Everyday Information").

#### Thinking Critically About Everyday Information: School Backpacks and Posture

A news report by MSNBC describes a study in which children were observed carrying school backpacks. The article states:

Thirteen children ages 8 and 9 walked about 1,310 feet without a backpack, and wearing packs weighing 9 and 13 pounds, while researchers filmed them with a high-speed camera. . . . The kids did not change their strides, the images showed. Instead, the youngsters bent forward more as they tried to counter the loads on their backs, and the heavier loads made them bend more, the study found. As they grew more tired, their heads went down, Orloff said.

Think about the following questions:

- Would you classify this as participant or nonparticipant observation? With awareness or without awareness?
- Could the answer to the previous question have influenced the children's behavior? If so, in what way?
- How could the dependent variables "stride" and "bend" be defined in terms that would permit quantitative measurement?
- Why do you believe the researcher used a high-speed camera instead of simply having human observers record observations?

SOURCE: Retrieved June 11, 2003, online at http://www.msnbc.com/news/922623.asp?0si=-

## **Reliability of Observations**

It is important to undertake a periodic check of the accuracy of your recorded observations by having another observer *independently* record his or her observations for the same time periods. These **independent observations** must not be influenced by the original observer's opinions or behavior. To accomplish this objective, it may be necessary to shield the two observers from each other. Even subtle recording movements on the part of one observer may be sufficient to bias the other observer. Having assured ourselves that the observations are independent, we can do a check for interobserver reliability.

#### Interobserver Agreement

We have noted several times that when different raters, judges, or observers are used to record data, it is important to determine whether the observations are objective and reliable. We try to assure this objectivity by carefully defining the behavior of interest and clearly specifying criteria for the occurrence or nonoccurrence of the behavior. Only when the behavior meets the criteria will it be counted as an observation. However, this procedure is not sufficient in itself. We must have evidence that our observations are objective and reliable, and that we have avoided observer bias, subjectivity, and observer drift. We simply cannot determine the accuracy of our data based upon the observations of a single observer. Therefore, a careful researcher will periodically use two or more observers simultaneously and then calculate a statistical measure to determine the degree of **interobserver agreement**. Several ways of calculating interobserver agreement will be described in the next section.

High interobserver agreement suggests that the behavior being observed is sufficiently well defined that we can generalize the results recorded by one observer to a population of observers. This then makes the behavioral phenomenon more meaningful to the individual investigator and to other investigators as well. Low interobserver reliability can cause problems. It could reduce the likelihood of finding an empirical relationship between the independent and the dependent variable. This would be unfortunate and wasteful of time and energy if, in fact, a relationship did exist. On the other hand, if an empirical relationship is found, low interobserver reliability would most likely diminish the confidence that one has in the firmness of the relationship. If we cannot obtain high interobserver agreement in spite of strong efforts to do so, then little confidence should be placed in the phenomenon because it may be impossible to detect systematic behavior of any kind or to assess the effects of any treatment. Reliability checks are expensive and time-consuming, but essential. Evaluation of interobserver agreement should be undertaken before the experiment begins and periodically thereafter. If agreement is either low or variable, then additional work is needed on defining behaviors, establishing criteria, and training observers. When agreement is low, a discussion should take place immediately after the interobserver agreement check to develop new rules and techniques that might improve reliability.

Let's briefly summarize the steps for maintaining observer reliability.

- 1. Establish objective criteria (decision rules) for determining whether the behavior did or did not occur.
- 2. Before you begin collecting data, conduct pilot testing to determine whether interobserver agreement is high with the established criteria.
- 3. If agreement is low or variable, additional work is needed. Reevaluate definitions or criteria. Behavior may not be well defined. Consider a training program for observers. Videotape can be very helpful.
- 4. If agreement is high, begin your study, but make periodic checks on observer reliability.

- 5. Periodic retraining may be necessary to avoid observer drift—that is, drifting away from established criteria or definition. Continue to assess interobserver agreement periodically.
- 6. If possible, use observers "blind" to the purposes of the study to prevent bias. If this is not possible, attempt to use a "blind" second observer when assessing interobserver agreement.

#### Measuring the Reliability of Observational Data

As we have seen, the researcher commonly has one of three different recording procedures from which to select: frequency of occurrence of a target behavior, duration of the occurrence of that behavior, and the occurrence versus the nonoccurrence of a behavioral event within a time interval. We will now consider some of the methods that are available to assess the reliability of observational data. More specifically, we will look at measures that involve percentage agreement among observers and correlational procedures modeled on the classical psychometric approach to reliability.

**Percentage Agreement Among Observers.** Let's suppose we are observing self-mutilating behavior among autistic children, and that we agree upon its definition. Using the occurrence/nonoccurrence procedure, we obtain the data shown in Table 6.3. Note that two different observers have independently recorded the presence or absence of self-mutilating behavior over four different observational intervals and five different sessions.

Table 6.3 Occurrence or Nonoccurrence Data Involving Two Independent Observers, Four Observational Intervals, and Five Sessions											
OBSERVATIONAL INTERVAL, OBSERVER A						0	BSERV		IAL II RVER		AL,
Session	1	2	3	4	Total	Session	1	2	3	4	Total
1	0	1	1	0	2	1	1	1	1	1	4
2	1	1	1	1	4	2	1	1	0	1	3
3	0	1	0	1	2	3	0	0	0	1	1
4	0	0	0	0	0	4	1	0	0	0	1
5	1	1	0	1	3	5	0	1	1	1	3

There are several methods of calculating the **percentage agreement among observers** for these data. One that is direct and easily understood simply involves dividing the session total of the observer with the smaller value by the corresponding session total of the observer with the larger value. Multiplying the resulting proportion by 100 yields a percentage agreement. To illustrate, on Session 1 Observer A's total was 2 and B's total was 4. Dividing 2 by 4 and multiplying by 100 yields a percentage agreement of 50%. Similarly, the percentages for

sessions 2, 3, 4, and 5 are, respectively,  $3/4 \times 100 = 75\%$ ,  $1/2 \times 100 = 50\%$ ,  $0/1 \times 100 = 0\%$ ,  $3/3 \times 100 = 100\%$ . The main limitation of this measure is that it is highly dependent on the rate at which the behavior is occurring during a given session. If the rate is either high or low, so that the target behavior either occurs or fails to occur during most or all observational intervals, the percentage agreement will be correspondingly high. Under these circumstances, it is also difficult to specify what constitutes chance levels of agreement.

A second percentage agreement statistic focuses on the percentage of session scores for which there is complete agreement between the two observers. Referring back to the total columns in Table 6.3, we see that the two observers have complete agreement only during session 5. Since there are five sessions, the percentage of agreement is  $1/5 \times 100 = 20\%$ . Because this percentage measure of reliability imposes a strict criterion of agreement (both totals must be identical) and does not utilize much of the data, it is not often used as a measure of interobserver reliability.

At times, your observations will attempt to classify a person into one category or another. For example, you may want to classify a child as aggressive or nonaggressive. Kappa is a statistic that will measure the proportion of agreement between observers and will correct for the level of chance agreement (Cohen, 1960). This is particularly important when one of the categories represents the "typical" behavior and occurs with a higher frequency.

Table 6.4 shows the frequency with which two observers agreed and disagreed across a total of 100 instances of behavior. Notice that in 50 instances both observer A and observer B agreed that the behavior was nonaggressive and that in 20 instances both agreed that the behavior was aggressive. The frequencies expected by chance are calculated by multiplying the frequency for the row times the frequency for the column and dividing by the total number of observations (*N*).

Table 6.4 Frequency of Agreement by Two Observers							
			<b>0</b> B S	ERVER A			
		CATEGORY	Nonaggressive	Aggressive			
OBSERVER B		Nonaggressive	$f_{\rm O} = 50$	f <sub>O</sub> = 10	60		
			$(f_{\rm C} = 42)$	$(f_{\rm C} = 18)$			
		Aggressive	$f_{\rm O} = 20$	$f_{\rm O} = 20$	40		
			$(f_{\rm C} = 28)$	$(f_{\rm C} = 12)$			
			70	30	N = 100		
Norm C C	1	1.0.0	1 1				

NOTE:  $f_{\rm O}$  = frequency observed;  $f_{\rm C}$  = frequency expected by chance

Kappa is calculated by using  $f_0$  and  $f_C$  on the diagonal where the categories match. Thus:

$$\kappa = \frac{f_{\rm O} - f_{\rm C}}{N - f_{\rm C}} = \frac{(50 + 20) - (42 + 12)}{100 - (42 + 12)} = \frac{70 - 54}{100 - 54} = \frac{16}{46} = 0.35$$

Notice that Kappa will equal a value of negative one (– 1.0) when there is absolutely no agreement, will equal a value of zero (0) when agreement is at chance level, and will equal a value of one (1.0) when agreement is complete and perfect. The Kappa value that represents acceptable reliability of observations depends on several factors, but researchers generally seek Kappa values greater than 0.75. Therefore, the above example illustrates a situation in which either the categories were not well defined or one (or both) of the observers was not doing a good job.

The Reliability Coefficient. The use of a measure of correlation—usually the Pearson product moment coefficient (r)—finds its origin in the psychometric tradition. The reliability of a test is expressed in terms of the size of the correlation coefficient. Although the Pearson r may vary from -1.00 to +1.00, it is rare that a negative **reliability coefficient** is found. For all practical purposes, we can assume that reliability coefficients vary between 0.00 and 1.00, with r = 0.00 meaning an absence of reliability and 1.00 meaning perfect reliability.

When using correlation to establish the reliability of observers, we regard the session total as a score. If two observers are in complete agreement, their totals for each observational session should be identical. Table 6.5 illustrates three different degrees of relationship between two observers in which N (the number of sessions) is 5. This is shown only for illustrative purposes, since N is not sufficiently large to establish the reliability with any given degree of confidence.

As we indicated, when observers are in complete agreement on their session totals, the correlation is 1.00. Generally, it is reasonable to assume that a high correlation means that both observers made the same or highly similar observations. On occasion, this may not be the case. It is possible to obtain a high or even a perfect correlation and still have observations that are not in agreement. This occurs because the correlation reflects only the relative position of paired observations and not the absolute values of these observations.

Table 6.5	Three Degre	Three Degrees of Correlation Between the Session Totals of Two Observers							
0	A negative correlation should rarely be found; if obtained, a serious review of observational procedures and recording techniques would be warranted.								
	<b>0</b> B S E	RVER	<b>O</b> B S E	RVER	<b>0</b> B S E	RVER			
Session	A	В	A	В	A	В			
1	2	4	2	3	2	3			
2	4	3	4	4	4	1			
3	2	1	2	1	2	1			
4	0	1	0	1	0	4			
5	3	3	3	3	3	3			
	r =	r = 0.58		r = 0.83		).68			

Imagine, for example, that two observers were simultaneously but independently rating a number of individuals on a scale of self-assertiveness. They obtain the results shown in Table 6.6.

Table 6.6 Hypothetical Data Showing Independent Ratings by Two Observers of Five Participants on Self-Assertiveness							
	<b>O</b> B S E R	VER A	OBSERVER B				
Participant	Rating	Rank	Rating	Rank			
1	18	3	22	3			
2	16	4	20	4			
3	20	2	24	2			
4	24	1	28	1			
5	14	5	18	5			

Note that the ranks of the ratings are in perfect agreement. The individual judged as highest in self-assertiveness by Observer A was also rated highest by Observer B. However, the ratings of Observer B were systematically four units higher than those of Observer A. As noted in Chapter 5, it is important to realize that reliability estimates reflect the relative position (rank order) of individual scores. Interrater reliability may be very high in making observations, yet it is possible for one rater to be consistently higher or lower in the total number of observations that are made. For this reason, measures of central tendency should accompany reliability ratings. If this difference went uncorrected and each observer were subsequently assigned to different experimental conditions, this difference between observers would be confounded with the independent variable. It would not be possible to separate the confounded effects of the independent variable from those of the observer. To eliminate this possibility of confounding, each

observer should be required to observe an equal number of times under each experimental condition. However, although the confounding would be corrected, the interobserver differences would add to the variability of the dependent measure. Consequently, the ability to detect differences among experimental conditions would be reduced. What this all means is that utmost care must be taken to assure interobserver reliability during all phases of the study.

The preceding example involved reliability of observations when the variable was quantitative in nature. The phi coefficient ( $\Phi$ ) is a correlation coefficient that can be used to measure the relationship when the variable is qualitative in nature—that is, on a nominal scale of measurement. You simply assign values of "0" to one category and values of "1" to the other category and then calculate the phi coefficient in the same manner that you would calculate a Pearson r. Notice that the sign (positive or negative) on the phi coefficient is not relevant because the assignment of 0s and 1s was arbitrary.

Thus, we see that observations are a common method of data collection in behavioral research but that special care is needed to ensure the quality of such observations. Depending on the specific nature of the observations, we can assess inter-observer reliability in different ways. In most cases, values for reliability coefficients vary from 0.00 to 1.00, with a higher value indicating greater reliability. There is no set standard for an acceptable level of reliability between observers; what is acceptable will depend to some extent on the variable being observed. Generally speaking, researchers should seek coefficients of .85 and higher.

## **Recordings by Equipment**

Although observational methods of data collection are useful and widespread, it is clear that there can be a degree of subjectivity in the process. It is also evident that some variables cannot be measured or are poorly measured via the human senses and that some form of equipment is needed. For example, most physiological measures (heart rate, blood pressure, sweat gland activity, brain electrical activity, hormone levels, chemical levels) require equipment.

Equipment often increases precision in data collection. However, it is important to check that all equipment is functioning properly and calibrated prior to data collection. If the data collection extends over some period of time, periodic checks of equipment accuracy are important.

#### **Public Records**

A wealth of information regarding human behavior is available in the public domain. This information includes census data, crime statistics, voting patterns, and national survey results. Using such data, a

researcher can describe human behavior and can ask questions regarding possible relationships among variables. This research strategy is often referred to as archival research because it relies on existing records. For example, is there a relationship between race and voting patterns? Public polling of people who had just voted in the 2000 national election showed that European Americans were more likely to vote Republican and African Americans were more likely to vote Democratic. As another example, a colleague of one of your authors has been investigating factors that relate to recidivism (repeat offending) in juvenile offenders. She received permission from the Department of Youth Services and the clinical psychologist who works with these juveniles to examine the criminal and psychological records of single offenders and multiple offenders. She is interested in whether juveniles diagnosed with psychological disorders are more likely to commit multiple offenses.

Although public records provide a wealth of information, they also have drawbacks. As a researcher, you did not collect the data. Therefore, you must investigate the methodology used to collect the data. Were sampling procedures involved? If so, were they adequate? Who collected the data? How did they collect the data? How have the data been stored? Can the accuracy of the data be trusted? These questions should be answered before you begin to explore and analyze data from public records.

# **Survey Methods**

**Survey** is a broad term that often includes interviews, questionnaires, and instruments or inventories. Let us make a few general comments regarding surveys in the broad sense. Several specific issues, such as the advantages and limitations of sampling techniques, will be detailed in the next chapter.

A survey is an attempt to estimate opinions, attitudes, and characteristics of a large number of individuals based on data collected from some of those individuals. The Kinsey reports in 1948 and 1953 provide well-known examples. Kinsey and his group interviewed more than 10,000 men and women regarding their sexual behavior and attitudes. Unfortunately, participants were not chosen randomly (each individual did not have an equal chance of being chosen). It is therefore debatable whether the findings can be generalized to the general population. Other researchers surveying sexual beliefs and practices have started with a sample of individuals representative of the population. However, only about 20% of those contacted were willing to share their beliefs and practices. Because of this low response rate, the resulting sample can hardly be regarded as representative of the population. Individuals who are willing to divulge intimate information are probably different in important ways from those who are not.

If survey results are to apply to a population, the sample chosen must be representative. When this principle is violated, serious problems can occur, and risk of error is high. For example, in the presidential election of 1936 between Alf Landon and Franklin Roosevelt, the *Literary Digest* conducted a poll concerning voting preferences. Based on the results of their poll, the *Literary Digest* predicted a Landon

(Republican) victory. As you know, the outcome of the election was Roosevelt by a landslide. This survey is a classic case of the unrepresentative sample. The magazine polled only those whose names appeared on lists of telephone subscribers and automobile owners. Because 1936 was a depression year, only wealthier people had cars and telephones—hardly a representative group. Affluent people tend to favor more conservative politics. Thus, if only the wealthier citizens had been allowed to vote, the Republican candidate would have won by a landslide.

Surveys, under different guises, have been used to obtain information on political opinions, consumer preferences, health care needs, abortion, and many other questions. The four most familiar are the U.S. Census, the Gallup Poll, the Roper Poll, and the Harris Poll. It is also quite common for news agencies and other organizations to conduct both phone surveys and Internet surveys.

#### Questionnaires

The **questionnaire** is more than simply a list of questions or forms to be completed. When properly constructed, a questionnaire can be used as a scientific instrument to obtain data from large numbers of individuals. Construction of a useful questionnaire that minimizes interfering problems requires experience, skill, thoughtfulness, and time. A major advantage of the questionnaire is that data can be obtained on large numbers of participants quickly and relatively inexpensively. Further, the sample can be very large and geographically representative. Often, anonymity can be easily maintained; that is, identifying information is not associated with the data. When constructed properly, a questionnaire provides data that can be organized easily, tabulated, and analyzed. Because of these apparent advantages, the use of the questionnaire is a popular method.

There are two broad classes of questionnaires: descriptive and analytical. Descriptive questionnaires are usually restricted to factual information, often biographical, which is usually accessible by other means. Job application forms and U.S. Census questionnaires are typically of this type. Analytical questionnaires deal more with information related to attitudes or opinions.

The results of a questionnaire are about as useful as the care and thought that went into its preparation and dissemination. Just as in normal social intercourse, the way questions are formulated and posed may present problems. They may be ambiguous; they may suggest the answer that the researcher "wants"; they may contain loaded words. Ambiguity is relatively easy to eliminate. A pilot project, limited to a small number of respondents, will usually uncover sources of ambiguity of which the researcher was unaware. These may then be corrected. Table 6.7 illustrates several examples of ambiguous and leading survey questions and also suggests improved versions of the questions.

Table 6.7 Examples of Poor Survey Questi	ons and Better Survey Questions.				
Poor Survey Question	BETTER SURVEY QUESTION				
How religious are you? (vague; not sure what is being asked)	How would you rate your level of spirituality?  Not at all Somewhat Very spiritual spiritual spiritual 0 1 2 3 4				
What do you think about smoking on campus? (vague; too many possible answers)	Do you believe that all buildings on campus should be designated as smoke-free?  Yes No				
How important is spirituality in your life?  Not at all Somewhat Very much so  1 2 3 4 5  (leading; suggests that spirituality is important)	What role does spirituality play in your life?  Not Somewhat Very important important  1 2 3 4 5				
What policy would best deal with the annoying and unhealthy effects of smoking on campus?  (a) smoking should be permitted anywhere (b) smoking only in designated areas of buildings (c) smoking only in outdoor areas (d) no smoking on campus (leading; suggests greater restrictions)	With regard to smoking on campus, which policy do you prefer?  (a) smoking should be permitted anywhere (b) smoking only in designated areas of buildings (c) smoking only in outdoor areas (d) no smoking on campus				

As much as we might wish it to be, completing questionnaires is not a neutral task, devoid of feelings and emotions. Often respondents are somewhat apprehensive about how they will appear in the researcher's eye. They want to look good and do well. Consequently, their responses may reflect their interpretations of the investigator's desires rather than their own beliefs, feelings, or opinions. This is referred to as demand characteristics. We will say more about this later. Obviously, questions should be stated in a neutral way and not in a way that suggests a particular response. A fundamental requirement is that the question should be answerable. If respondents are given answers from which to choose, the options should be clear and independent. Also, different results can occur when open-ended or closed-ended questions are used. In some cases, the questionnaire is sensitive to position effects. Respondents are more likely to skip items placed toward the end of a questionnaire, and the answers are also slightly different when answered.

More attention has been given to response bias than to other sources of possible bias and contamination. As we noted earlier, results can be markedly affected by the sample on which they are based. The problem of sampling bias is compounded in mailed surveys because of the low return rates. The actual sample on which the data analyses are based is generally a subsample of the original sample. Low returns make it difficult to assess the representativeness of the final sample. It is safe to assume that it is biased and that those who participated in the survey are different in some way from those who did not. How important is this difference? It may be considerable, or it may be trivial. Because its importance

cannot be assessed, any generalizations based on low returns must be restricted. For this reason, it is important to know the return rate on survey research. Unfortunately, some studies fail to provide this information. Other things being equal, the higher the return rate, the better the survey.

A number of factors affect return/response rates. Some are quite costly, so that economic factors must be balanced against the greater generality permitted by higher rates of return. Methods to increase return rate include follow-up contacts, general delivery and pickup, use of closed-ended rather than open-ended questions wherever possible, use of rewards for participation, and limiting the length of time needed to complete the survey.

#### **Instruments and Inventories**

Instruments and inventories are questionnaires that have stood the test of time. That is, they were designed to measure particular attributes and have been demonstrated to do so with validity and reliability. Examples include personality tests, aptitude tests, and achievement tests. Personality tests measure some state or trait of an individual. Examples include the Minnesota Multiphasic Personality Inventory (MMPI), Beck Depression Inventory (BDI), California Psychological Inventory (CPI), and the Sixteen Personality Factors Questionnaire (16PF). Aptitude tests measure some skill or ability. Examples include the Stanford–Binet Intelligence Scale, the Wechsler Adult Intelligence Scale (WAIS-III), the Wechsler Intelligence Scale for Children (WISC-III), and the Graduate Record Examination (GRE). Achievement tests measure competence in a particular area. Examples include the Stanford Achievement tests that students take as they progress through K–12 grades in school; state licensing exams for teachers, counselors, lawyers, physicians and other professionals; and the major field achievement test that psychology majors at some universities take just prior to graduation.

If you consider a research project in which a questionnaire might be used, it would be wise to determine whether an instrument or inventory already exists to measure the variable of interest. Don't reinvent the wheel. If someone else has already invested the time and effort to develop a measure with known validity and reliability, use it. One of the characteristics of science is that we make information public and continue to build upon what others have done.

#### Interviews

The **interview** may be regarded either as an alternative to other survey methods or as a supplementary source of information. Although it is more costly in both time and money than the questionnaire, it is also more flexible. Additional information over and above initial plans can be readily obtained and ambiguity and misunderstanding eliminated immediately.

One of the greatest strengths of the interview—direct verbal communication—is also a source of weakness because variability is so common in social interactions. For an interview to be successful, rapport is generally required. It is most readily established when the interviewer is nonjudgmental, supportive, and understanding. However, these very characteristics lead to variability in social interaction among those interviewed. We could achieve sufficient control over social interactions so that the interviews are more homogeneous. However, this would inevitably lead to a sterile interview situation. This, in turn, would result in less rapport, which, we have noted, is important for a good interview.

Other problems beset the interview, especially when there is more than one interviewer. Different interviewers may vary in the way they ask questions or interpret responses, or in the way respondents react to them. Interviewer differences are common. How do we assess the comparability of different interviewers? If you reflect a moment, you'll realize that the situation is similar to using several raters in noninterview settings and determining the interrater reliability. In the present case, we are asking whether there is inter-interviewer reliability.

One way to achieve greater inter-interviewer reliability is to standardize the interview procedures. While this standardization increases the interview reliability, it decreases its flexibility. Because of these weaknesses, the interview might best be reserved as an exploratory method to generate ideas and hypotheses that can later be tested by the use of other methods.

Table 6.8 summarizes some general tips for the development of effective surveys.

### Table 6.8 Tips for Effective Surveys

- 1. Avoid ambiguous questions.
- 2. Avoid leading questions.
- 3. Avoid lengthy surveys or those that require lengthy written responses.
- 4. Consider how initial questions might influence answers to subsequent questions.
- 5. Think carefully about your sampling technique to avoid a biased sample.
- **6.** Seek to achieve the highest response rate possible.
- 7. Standardize administration procedures.
- 8. Guarantee anonymity (or confidentiality at a minimum).
- 9. Seek measures of reliability.
- **10.** Assess validity.

## Laboratory vs. Field Research

Data can be collected either in the laboratory or in the field. **Laboratory research** takes place in an environment designed by the researcher, whereas **field research** takes place in an environment not

designed by the researcher. Most of the experimental and nonexperimental designs that will be discussed in subsequent chapters can be carried out either in the laboratory or in the field. However, it is true that experimental designs are more likely to be implemented in the laboratory and nonexperimental designs are more likely to be implemented in the field.

As we will see, high levels of control characterize experimental designs. The experimenter controls the assignment of participants to conditions, manipulates the levels of the independent variable, and seeks to eliminate extraneous variables that may affect the dependent variable. These objectives are much easier to accomplish in a laboratory environment than in a natural setting. Thus, the researcher is often more confident in drawing strong cause–effect conclusions when the experiment is conducted in the laboratory.

Field research has its own advantages. Research conducted in a natural setting is more likely to involve natural behavior, and thus the findings are more likely to generalize to the "real world." If we are interested in generalizing or applying our experimental findings to practical problems or to social issues, then this external validity is essential. Further, if we are interested in understanding and predicting human or nonhuman behavior in situations other than those studied in the laboratory, then our research methods should be more representative of these situations. Some examples will clarify the distinction between laboratory and field research.

Social psychologists have used field research frequently and successfully; they also, of course, use laboratory research. An excellent film, *Bystander Intervention: When Will People Help?* depicts both types. Bystander intervention is one area in which both field and laboratory experiments have been conducted. In the typical field experiment on bystander intervention, a person suddenly collapses in full view of other pedestrians. The researchers are usually interested in the number of people who come to the aid of the distressed person and how long it takes for the person to receive help. A number of factors, such as age, race, gender, appearance of the individual, location in the city, and time of day, may be systematically varied in order to assess their effects on helping behavior. Observers stationed nearby can unobtrusively collect different kinds of data, such as the proportion who offer aid, the time it takes before aid is offered, who offers aid, and so forth. Follow-up questionnaires or interviews may be used to pinpoint more precisely the differences between those who offered aid and those who did not.

Another example of a field experiment is the "lost letter routine." The investigator writes letters addressed to himself or herself, complete with necessary postage. The letters are then dropped ("lost") at various locations. The dependent variable is the number of letters returned. The independent variable may be any number of different factors. For example, the address could be a political party, a religious group, a political candidate, a government agency, or a particular section of the city. Individuals finding a "lost letter" must decide what to do with it—return it, open it, ignore it, or discard it. Dependent variable

measures noted in this and the previous paragraph are referred to as **unobtrusive measures** because individuals are unaware that they are being measured.

During the research design phase, you must consider whether data collection will occur in the laboratory or in the field. In most cases, the research question itself dictates the appropriate setting. The more important point is to be aware of the advantages and limitations of each setting. The laboratory affords a high level of control and leads to more powerful and confident conclusions regarding the relationships among the variables under study. The field affords a natural setting, permits research not possible in the laboratory, and leads to conclusions that are more likely to generalize to the real world.

## **Case Analysis**

One of your authors, along with several undergraduate research assistants, is interested in the types of teaching techniques that professors use in the classroom. One specific question concerns the prevalence of lecture in the university classroom and whether male and female instructors use this teaching technique to different extents. Rather than rely on self-reports from professors, this study attempts to answer the question by using observational techniques.

Two student researchers randomly sample 20 male and 20 female professors to observe. They inform the professors of the purpose of their study and obtain consent to sit in on one class for each of the professors. The student observers coordinate their schedules and decide on the class to be observed for each professor. For each class, they find two seats near the back of the room and record, at 10-minute intervals, whether the professor is lecturing. If the class begins at 8:00 and ends at 9:15, the student observers record either "yes" or "no" at 8:10, 8:20, 8:30, 8:40, 8:50, 9:00, and 9:10. After observing all 40 professors, they total the number of times that male professors were lecturing and compare that to the number of times that female professors were lecturing.

#### Critical Thinking Questions

- 1. Were the two students participant observers or nonparticipant observers? Which technique do you believe would be best in this situation?
- 2. How did the observers schedule their observations (pick the class, day, time)? Can you think of a better way to do this?
- 3. Which specific technique was used to record the lecturing behavior? Do you believe that this was the best method to answer the original research question? How would you use the duration method instead?
- 4. What should the two students do to ensure that their observations are independent?
- 5. What statistic should they use to measure the reliability of their observations?

6. In what other ways could the design of this research be improved?

## **General Summary**

Behavioral research often uses dependent variables that rely on identification by the researchers. Several issues must be considered during the design phase of the research. Will the researcher engage in participant or nonparticipant observation? Participant observation occurs when the observer is an integral part of the environment to be observed. This allows the observer to be closer to the behaviors being observed, but may inadvertently affect those behaviors. If this is a serious risk, the researcher should consider nonparticipant observation, remaining "outside" the environment to be observed. After deciding on the type of observation technique, the researcher must also decide on the schedule for observations, the definition of the behavior to be observed, and the aspect of the behavior to be observed. Sometimes it is appropriate to use a frequency method, in which the observer records the number of times that a defined behavior occurs. Sometimes it is appropriate to use a duration method, in which the observer records how long a defined behavior lasts. At other times it is appropriate to use an interval method, in which the observer determines whether or not a particular behavior occurs at specified intervals of time.

Observers rely on sensory systems and perceptual processing to record observations. Psychological research shows that a multitude of factors influence our sensation and perception. This argues for the value of multiple observers. A high measure of reliability across multiple observers enhances the objective nature of our research. A low measure of reliability indicates a degree of idiosyncrasy in the observations and reduces our confidence in the objectivity of our research.

In addition to observations made by researchers, other methods of data collection include recordings by equipment, public records, and survey methods, including questionnaires, instruments/inventories, and interviews. Each method has advantages and disadvantages. Finally, all methods of data collection can occur either in a laboratory or in a field setting. A laboratory affords more control over the environment, and the field affords a more natural setting. Whichever method of data collection is used in whichever setting, the researcher must decide which members of the population will be observed. The next chapter discusses these sampling techniques.

# **Detailed Summary**

- 1. Observational procedures represent a methodology that is critical to the field of behavioral science. Observations can be appropriate in natural, applied, and laboratory settings.
- 2. Because observations are based on our perceptions, the same event occurring in the environment will be perceived differently by different people.
- 3. It is essential that observations be accurate, objective, and reliable.

- 4. Participant observers conduct their observations "from the inside"; that is, the researcher is an integral part of the environment being observed. Nonparticipant observers conduct their observations "from the outside"; the researcher does not interact with those being observed.
- 5. When making observations, investigators must make decisions regarding frequency, duration, and time of day for scheduling observations. These decisions depend on the purposes of the observations.
- 6. It is important to define the dependent variable (behavior) in terms of specific observable responses and to specify clearly the criteria for judging when the behavior has occurred. This is an important step if we are to ensure that different observers will make similar observations.
- 7. With the frequency method of observation, the observer simply counts the number of occurrences of the behavior of interest in a given interval of time. With the duration method, the observer records the beginning and the end of a particular behavior. With the interval method, the observation period is broken into equal intervals, the size of which varies with the particular observations of interest. Behavior is recorded as occurring or not occurring in each interval.
- 8. It is important to assess the reliability of observations by having another observer independently record his or her observations for the same time periods. Depending on the nature of the data, reliability can be measured by percent agreement, a Kappa statistic, a Pearson *r* statistic, or a phi coefficient.
- In addition to human observers, we can use equipment, public records, surveys, questionnaires, instruments/inventories, and interviews to collect data. Each method has advantages and disadvantages that should be carefully considered.
- 10. Data can be collected either in the laboratory or in the field. Laboratory research takes place in an environment designed by the researcher, whereas field research takes place in an environment not designed by the researcher. Laboratory research has the advantage of control, and field research has the advantage of natural, real-world behavior.

# **Key Terms**

duration method of observation field research frequency method of observation independent observation instrument inter-observer agreement interval method of observation interview inventory
laboratory research
nonparticipant observation
observer drift
participant observation
percentage agreement among observers
questionnaire
reliability coefficient
survey
unobtrusive measures

## **Review Questions / Exercises**

Several years ago, a group of students at the University of Central Arkansas conducted a study in which they observed the rate at which cars failed to stop at a campus stop sign and recorded whether the car had a student parking decal or a faculty/staff parking decal. Use the above study to answer questions 1–7.

- Which method of observation would be best (participant or nonparticipant, aware or unaware)?
   Justify your answer.
- 2. How would you schedule observations?
- 3. Define the categories of behavior that you would observe.
- 4. Describe how you would optimize and measure the reliability of observations, including the use of independent observers and calculation of interobserver agreement.
- 5. Describe how you could use equipment for observations rather than human observers. What are the advantages and disadvantages?
- 6. Describe how you might use public records to answer the same research question. What might be some limitations of this approach?
- 7. Describe how you might use a survey method to answer the same research question. What might be some limitations of this approach?