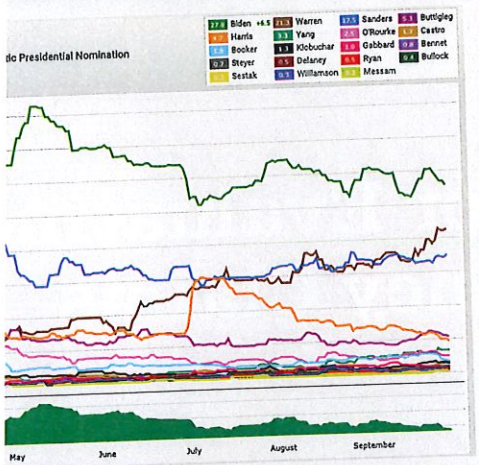
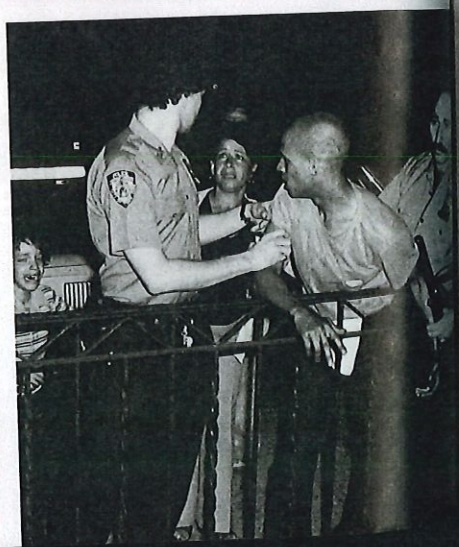


irds of Adults Have Had an
e Childhood Experience
GH, 2017

Chabot Polling at 50%

REALCLEARPOLITICS, 2016



6

Surveys and Observations: Describing What People Do

YOU SHOULD BE ABLE to identify the three statements that open this chapter as single-variable frequency claims. Each claim is based on data from one self-reported variable: the rated quality of a restaurant, support for a congressional candidate, or people's history of adverse childhood experiences (such as abuse or neglect). Where do the data for such claims come from? This chapter focuses on the construct validity of surveys and polls, in which researchers ask people questions, as well as observational studies, in which researchers watch the behavior of people or animals, often without asking them questions at all. Researchers use surveys, polls, and observations to measure variables for any type of claim. However, in this chapter and the next, many of the examples focus on how surveys and observations are used to measure one variable at a time for frequency claims.

CONSTRUCT VALIDITY OF SURVEYS AND POLLS

Researchers use surveys and polls to ask people questions over the Internet, in door-to-door interviews, or on the phone. You may have been asked to take surveys in various situations. Perhaps after you

LEARNING OBJECTIVES

A year from now, you should still be able to:

1. Explain how carefully prepared questions improve the construct validity of a poll or survey.
2. Describe how researchers can make observations with good construct validity.

purchased an item from an Internet retailer, you got an e-mail asking you to post a review. While you were reading news online, maybe a survey popped up. A polling organization such as Gallup may have called your cell phone.

The word *survey* is often used when people are asked about a consumer product, whereas the word *poll* is used when people are asked about their social or political opinions. However, in this book, **survey** and **poll** both mean the same thing: a method of posing questions to people online, in personal interviews, or in written questionnaires.

How much can you learn about a phenomenon just by asking people questions? It depends on how well you ask. As you will learn, researchers who develop their questions carefully can support frequency, association, or causal claims that have excellent construct validity.

Choosing Question Formats

Survey questions can follow several formats. Researchers may ask **open-ended questions** that allow respondents to answer any way they like. They might ask people to name the public figure they admire the most. Departing overnight guests might be asked to comment on their experience at a hotel. People's various responses to these open-ended questions provide researchers with spontaneous, rich information. The drawback is that the responses must be coded and categorized, a process that is difficult and time-consuming. In the interest of efficiency, researchers in psychology often restrict the answers people can give.

One specific way to ask survey questions uses **forced-choice questions**, in which people give their opinion by picking the best of two or more options. Forced-choice questions are often used in political polls, such as asking: If the Ohio congressional election were held today, would you vote for the Republican Steve Chabot? Or the Democrat Aftab Pureval?

Forced-choice questions are also used to measure personality. An example comes from the Narcissistic Personality Inventory (NPI; Raskin & Terry, 1988). This instrument asks people to choose one statement from each of 40 pairs of items, such as the following:

1. ____ I really like to be the center of attention.
   ____ It makes me uncomfortable to be the center of attention.
2. ____ I am going to be a great person.
   ____ I hope I am going to be successful.

To score a survey like this, the researcher adds up the number of times people choose the "narcissistic" response over the "non-narcissistic" one (in the pairs above, the narcissistic response is the first option).

Simple yes/no questions are also considered a forced-choice format. The Adverse Childhood Experiences (ACE) study asks about early history of violence, abuse, and neglect with simple yes/no questions like those depicted in **Figure 6.1** (Merrick et al., 2018).

QUESTION 1 OF 10

Before your 18th birthday, did a parent or other adult in the household often or very often . . .

swear at you, insult you, put you down, or humiliate you?
or
act in a way that made you afraid that you might be physically hurt?

YES | NO

QUESTION 2 OF 10

Before your 18th birthday, did a parent or other adult in the household often or very often . . .

push, grab, slap, or throw something at you?
or
hit you so hard that you had marks or were injured?

YES | NO

FIGURE 6.1
ACE Questions.
Questions from an online version of the Adverse Childhood Experiences (ACE) survey, which asks about early experiences with violence, abuse, and neglect.

In another question format, people are presented with a statement and are asked to use a rating scale to indicate their degree of agreement. When such a scale contains more than one item and each response value is labeled with the specific terms *strongly agree, agree, neither agree nor disagree, disagree,* and *strongly disagree,* it is often called a **Likert scale** (Likert, 1932). If it does not follow this format exactly (e.g., if it has only one item, or if its response labels differ from the original Likert labels), it may be called a *Likert-type scale.* Here is one of the ten items from a commonly used measure called the Rosenberg Self-Esteem Scale (Rosenberg, 1965), which uses a Likert scale:

I am able to do things as well as most other people.

1	2	3	4	5
Strongly disagree				Strongly agree

Instead of degree of agreement, respondents might be asked to rate a target object using a numeric scale that is anchored with adjectives; this is called a **semantic differential format**. For example, on the Internet site RateMyProfessors.com, students assign ratings to a professor using the following adjective phrases:

Overall Quality:

	1	2	3	4	5	
Profs get F's too						A real gem

2
restaurant rating on the

gs of products and services are
frequency claims. Is a five-star
d indicator of a restaurant's

Level of Difficulty:						
Show up and pass	1	2	3	4	5	Hardest thing I've ever done

The five-star rating format that Internet rating sites (like Yelp) use is another example of this technique (**Figure 6.2**). Generally one star means "poor" or (on Yelp) "Eek! Methinks not," and five stars means "outstanding" or even "Woohoo! As good as it gets!"

There are other question types, of course, and researchers might combine formats on a single survey. The point is that the format of a question (open-ended, forced-choice, or Likert scale) does not make or break its construct validity. The way the questions are worded and the order in which they appear are much more important.

Writing Well-Worded Questions

When you interrogate a survey result, your first question is about construct validity: How well was that variable measured? The way a question is worded and presented in a survey can make a tremendous difference in how people answer. It is crucial that each question be clear and straightforward. Poll and survey creators work to ensure that the wording and order of the questions do not influence respondents' answers.

QUESTION WORDING MATTERS

An example of the way question wording can affect responses comes from survey research on a random sample of Delaware voters (Wilson & Brewer, 2016). The poll asked about people's support for voter identification laws, which require voters to show a photo ID before casting a ballot. Participants heard one of several different versions of the question. **Table 6.1** presents the results.

As you can see, different versions of the question led to different results. The second and third wordings might be called **leading questions**, because their wording leads people to a particular response. The results show that the wording matters. When people answer questions that suggest a particular viewpoint, at least some people change their answers.

In general, if the intention of a survey is to capture respondents' true opinions, the survey writers might attempt to word every question neutrally, avoiding potentially emotional terms. When researchers want to measure how much the wording matters for their topic, they word each question more than one way. If the results are the same regardless of the wording, they can conclude that question wording does not affect people's responses to that particular topic. If

TABLE 6.1

Question wording matters

A random sample of Delaware voters were randomly assigned to one of three versions of a voter ID question.

WHEN THE QUESTION WAS WORDED LIKE THIS:	DELAWARE VOTERS' SUPPORT FOR VOTER ID LAWS WAS:
What is your opinion? Do you strongly favor, mostly favor, mostly oppose, or strongly oppose voter ID laws?	79%
Opponents of voter ID laws argue that they will prevent people who are eligible to vote from voting. What is your opinion? Do you strongly favor, mostly favor, mostly oppose, or strongly oppose voter ID laws?	69%
Opponents of voter ID laws argue that they will prevent people who are eligible to vote from voting, and that the laws will affect African American voters especially hard. What is your opinion? Do you strongly favor, mostly favor, mostly oppose, or strongly oppose voter ID laws?	61%

the results differ, then they may need to report the results separately for each version of the question.

DOUBLE-BARRELED QUESTIONS

The wording of a question is sometimes so complicated that respondents have trouble answering in a way that accurately reflects their opinions. In a survey, it is always best to ask a simple question. When people understand the question, they can give a clear, direct, and meaningful answer, but sometimes survey writers forget this basic guideline. For example, an online survey from the National Rifle Association asked this question:

> Do you agree that the Second Amendment to our United States Constitution guarantees your individual right to own a gun and that the Second Amendment is just as important as your other Constitutional rights?
>
> ____ Support
> ____ Oppose
> ____ No opinion

This is called a **double-barreled question**; it asks two questions in one. Double-barreled questions have poor construct validity because people might be responding to the first half of the question, the second half, or both. Therefore, the

item could be measuring the first construct, the second construct, or both. Careful researchers would have asked each question separately:

Do you agree that the Second Amendment guarantees your individual right to own a gun?

____ Support
____ Oppose
____ No opinion

Do you agree that the Second Amendment is just as important as your other Constitutional rights?

____ Support
____ Oppose
____ No opinion

NEGATIVE WORDING

Negatively worded questions are another way survey items can be unnecessarily complicated. Whenever a question contains negative phrasing, it can cause confusion, thereby reducing the construct validity of a survey or poll (Schwarz & Oyserman, 2001).

A classic example comes from a survey on Holocaust denial, which found that 20% of Americans denied that the Nazi Holocaust ever happened. In the months that followed the publication of this survey's results, writers and journalists criticized and analyzed the "intensely disturbing" news (Kagay, 1994).

Upon further investigation, the Roper polling organization reported that the people in the original telephone poll had been asked, "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" Think for a minute about how you would answer that question. If you wanted to convey the opinion that the Holocaust did happen, you would have to say, "It's impossible that it never happened." In order to give your opinion about the Holocaust accurately, you must also be able to unpack the double negative of *"impossible"* and *"never."* So instead of measuring people's beliefs, the question may be measuring people's working memory or their motivation to pay attention.

We know that this negatively worded question may have affected people's responses because the same polling organization repeated the survey less than a year later, asking the question more clearly: "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" This time, only 1% responded that the Holocaust might not have happened, 8% did not know, and 91% said they were certain it happened (Kagay, 1994). This later result, as well as other polls reflecting similarly low levels of Holocaust denial, indicates that the question as it was originally worded had poor construct validity: It probably did not measure people's true beliefs.

Sometimes even one negative word can make a question difficult to answer. For example, consider the following question:

Abortion should never be restricted.

1	2	3	4	5
Disagree				Agree

To answer this question, those who oppose abortion must think in the double negative ("I *disagree* that abortion should *never* be restricted"), while those who support abortion rights would be able to answer more easily ("I agree—abortion should never be restricted").

When possible, negative wording should be avoided, but researchers sometimes ask questions both ways, like this:

Abortion should never be restricted.

1	2	3	4	5
Disagree				Agree

I favor strong restrictions on abortion.

1	2	3	4	5
Disagree				Agree

After asking the question both ways, the researchers can study the items' internal consistency (using Cronbach's alpha) to see whether people respond similarly to both questions. (In this case, agreement with the first item should correlate with disagreement with the second item.) Like double-barreled questions, negatively worded ones can reduce construct validity because they might capture people's ability to figure out the question rather than their true opinions.

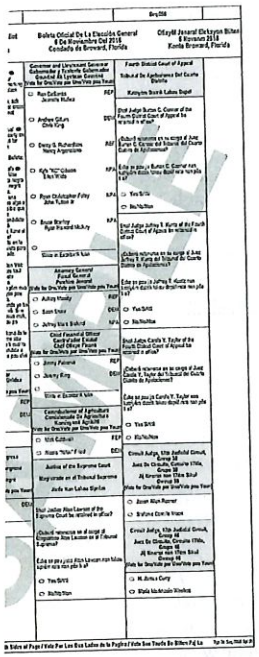
« For more on Cronbach's alpha, see Chapter 5, pp. 131–132.

QUESTION ORDER

The order in which questions are asked can also affect the responses to a survey. We might safely assume that respondents' answers to the first question on a survey is unaffected by any other questions, but the earlier questions sometimes change the way respondents understand and answer the later questions. For example, a question on a parenting survey such as "How often do your children play?" would have different meanings if the previous questions had been about sports versus music versus daily activities.

Consider this example: Political opinion researcher David Wilson and his colleagues asked people whether they supported affirmative action for different groups (Wilson et al., 2008). Half the participants were asked two forced-choice questions in this order:

1. Do you generally favor or oppose affirmative action programs for women?
2. Do you generally favor or oppose affirmative action for racial minorities?



6.3
on order matters.

ire up to 5% more likely to vote
irst person they see on a ballot
. Krosnick, 1998). Because of this
ı, a Florida court ruled in 2019 that
y in power does not have the right
s own candidates first.

The other half were asked the same two questions, but in the opposite order:

1. Do you generally favor or oppose affirmative action for racial minorities?
2. Do you generally favor or oppose affirmative action programs for women?

Wilson found that Whites reported more support for affirmative action for minorities when they had first been asked about affirmative action for women. Presumably, most Whites support affirmative action for women more than they do for minorities. To appear consistent, they might feel obligated to express support for affirmative action for racial minorities if they have just indicated their support for affirmative action for women (**Figure 6.3**).

The most direct way to control for the effect of question order is to prepare different versions of a survey, with the questions in different sequences. If the results for the first order differ from the results for the second order, researchers can report each set of results separately.

Encouraging Accurate Responses

Careful researchers pay attention to how they word and order their survey questions. But what about the people who answer them? Overall, people can give meaningful responses to many kinds of questions (Paulhus & Vazire, 2007; Schwarz & Oyserman, 2001). In certain situations, people might give inaccurate answers because they don't make an effort to think about each question, because they want to look good, or because they are simply unable to report accurately about their own motivations and memories.

PEOPLE CAN GIVE MEANINGFUL RESPONSES

Many students start out skeptical that people's self-reports can be trusted. But in fact self-reports can be ideal. People are able to report their own gender identity, happiness, income, ethnicity, and so on; there is no need to use expensive or difficult measures to collect such information. People can accurately report on things they did or that happened to them. More important, self-reports often provide the only meaningful information you can get. Diener and his colleagues were specifically interested in subjective well-being (see Chapter 5), so it made sense to ask participants to self-report on aspects of their life satisfaction (Diener et al., 1985). Who else but you knows how happy you feel?

In some cases, self-reports might be the only option. For example, researchers who study dreaming can monitor brain activity to identify *when* someone is dreaming, but they need to use self-reports to find out the *content* of the person's dreams. Other traits are not very observable, such as how anxious somebody is feeling or whether someone has been a victim of violence. Therefore, it is meaningful and effective to ask people to self-report on their own experiences (Becker-Blease & Freyd, 2006; Vazire & Carlson, 2011).

SOMETIMES PEOPLE USE SHORTCUTS

At times, however, self-reports are imperfect. **Response sets**, also known as *non-differentiation*, are a type of shortcut people can take when answering survey questions. Although response sets do not cause many problems for answering a single, stand-alone item, people might adopt a consistent way of answering all the questions—especially toward the end of a long questionnaire (Lelkes et al., 2012). Rather than thinking carefully about each question, people might answer all of them positively, negatively, or neutrally. Response sets weaken construct validity because these survey respondents are not saying what they really think.

One potential response set is **acquiescence**, or *yea-saying*. This occurs when people say "yes" or "strongly agree" to every item instead of thinking carefully about each one. For example, a respondent might answer "5" to every item on Diener's scale of subjective well-being—not because the respondent is happy, but because that person is using a yea-saying shortcut (**Figure 6.4**). People apparently have a bias to agree with (say "yes" to) any item—no matter what it states (Krosnick, 1999). Acquiescence can threaten construct validity because instead of measuring the construct of true feelings of well-being, the survey could be measuring the tendency to agree or the lack of motivation to think carefully.

How can researchers tell the difference between a respondent who is yea-saying and one who really does agree with all the items? The most common way is by including reverse-worded items. Diener might have changed the wording of some items to mean their opposite; for instance, "If I had my life to live over, I'd change almost *everything*." One benefit is that reverse-worded items might slow people down so they answer more carefully. (Before computing a scale average for each person, the researchers rescore only the reverse-worded items such that, for example, "strongly disagree" becomes a 5 and "strongly agree" becomes a 1.) The scale with reverse-worded items would have more construct validity because high or low averages would be measuring true happiness or unhappiness, instead of acquiescence.

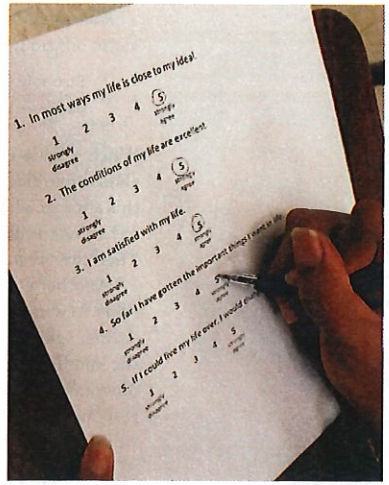


FIGURE 6.4
Response sets.
When people use an acquiescent response set, they agree with almost every question or statement. It can be hard to know whether they really mean it or whether they're just using a shortcut to respond to the questions.

A drawback of reverse-wording is that sometimes it results in negatively worded items, which are more difficult to answer.

Another specific response set is **fence sitting**—playing it safe by answering in the middle of the scale, especially when survey items are controversial. People might also answer in the middle (or say "I don't know") when a question is confusing or unclear. Fence sitters can weaken a survey's construct validity when middle-of-the-road scores suggest that some responders don't have an opinion, though they actually do. Of course, some people honestly may have no opinion on the questions; in that case, they choose the middle option for a valid reason. It can be difficult to distinguish those who are unwilling to take a side from those who are truly ambivalent.

Researchers may try to jostle people out of this tendency. One approach is to take away the neutral option. Compare these two formats:

Race relations are going well in this country.

○	○	○	○	○
Strongly disagree				Strongly agree

Race relations are going well in this country.

○	○	○	○
Strongly disagree			Strongly agree

When a scale contains an even number of response options, the person has to choose one side or the other because there is no neutral choice. The drawback of this approach is that sometimes people really do not have an opinion or an answer, so having to choose a side is an invalid representation of their truly neutral stance. Therefore, researchers must carefully consider which format is best.

Another common way to get people off the fence is to use forced-choice questions, in which people must pick one of two answers. Although this reduces fence sitting, again it can frustrate people who feel their own opinion is somewhere in the middle of the two options. In some telephone surveys, interviewers will write down a response of "I don't know" or "No opinion" if a person volunteers that response. Thus, more people get off the fence, but truly ambivalent people can also validly report their neutral opinions.

TRYING TO LOOK GOOD

Most of us want to look good in the eyes of others, but when survey respondents give answers that make them look better than they really are, these responses decrease the survey's construct validity. This phenomenon is known as **socially desirable responding**, or **faking good**. The idea is that because respondents are embarrassed, shy, or worried about giving an unpopular opinion, they will not tell the truth on a survey or other self-report measure. A similar, but less common, phenomenon is called **faking bad**.

To avoid socially desirable responding, a researcher might ensure that the participants know their responses are anonymous—perhaps by conducting the survey online, or in the case of an in-person interview, reminding people of their anonymity right before asking sensitive questions (Schwarz & Oyserman, 2001). However, anonymity may not be a perfect solution. Anonymous respondents may treat surveys less seriously. In one study, anonymous respondents were more likely to start using response sets in long surveys. In addition, anonymous people were less likely to accurately report a simple behavior, such as how many candies they had just eaten, which suggests they were paying less attention to details (Lelkes et al., 2012).

One way to minimize this problem is to include special survey items that identify socially desirable responders with target items like these (Crowne & Marlowe, 1960):

> My table manners at home are as good as when I eat out in a restaurant.
> I don't find it particularly difficult to get along with loud-mouthed, obnoxious people.

If people agree with many such items, researchers may discard that individual's data from the final set, under suspicion that they are exaggerating on the other survey items or not paying close attention in general.

Researchers can also ask people's friends to rate them. When it comes to domains where we want to look good (e.g., on how rude or how smart we are), others know us better than we know ourselves (Vazire & Carlson, 2011). Thus, researchers might be better off asking people's friends to rate them on traits that are observable but desirable.

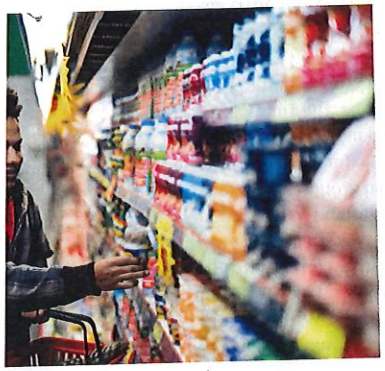
Finally, researchers may use computerized measures to evaluate people's implicit opinions about sensitive topics. One widely used test, the Implicit Association Test, asks people to respond quickly to positive and negative words on the right and left of a computer screen (Greenwald et al., 2003). Intermixed with the positive and negative words are instances of different social groups, such as males and females, Blacks and Whites, or Middle Eastern females and White females. People respond to all possible pairs. For example, they respond to positive words with Black faces, negative words with White faces, negative words with Black faces, and positive words with White faces. When people respond more efficiently to the White-positive/Black-negative combination than to the White-negative/Black-positive combination, researchers infer that the person may hold negative attitudes on an implicit, or unconscious, level (Jost, 2019).

SELF-REPORTING "MORE THAN THEY CAN KNOW"

In general, people are *capable* of reporting accurately on their own feelings, thoughts, and actions. Everyone knows their opinions better than anyone else

does. Only *I* know my level of support for a political candidate. Only *the patrons* know how much they liked that restaurant. In certain cases, however, self-reports can be inaccurate, especially when people are asked to describe *why* they are thinking, behaving, or feeling the way they do. When asked, most people willingly provide an explanation or an opinion to a researcher, but sometimes they unintentionally give inaccurate responses.

Psychologists Richard Nisbett and Timothy Wilson (1977) conducted a set of studies to demonstrate this phenomenon. In one study, they put six pairs of nylon stockings on a table and asked female shoppers in a store to tell them which of the stockings they preferred. As it turned out, almost everyone selected the last pair on the right. The reason for this preference was something of a mystery—especially since all the stockings were exactly the same! (Researchers have speculated people are biased toward the last item they evaluate.) Next, the researchers asked each woman why she selected the pair she did. Every participant reported that she selected the pair on the right for its excellent quality. Even when the researchers suggested they might have chosen the pair because it was on the far right side of the table, the women insisted they made their choices based on the quality of the stockings. In other words, the women easily formulated answers for the researchers, but their answers had nothing to do with the real reason they selected the one pair of stockings (**Figure 6.5**). Moreover, the women did not seem to be aware they were inventing a justification for their preference. They gave a sincere, reasonable response—one that just happened to be wrong. Therefore, researchers cannot assume the reasons people give for their own behavior are their actual reasons. People may not be able to accurately explain *why* they acted as they did.



: 6.5
curacy of self-reports.
:k this shopper why he chooses one of these
e will probably give you a reasonable answer. But
: answer represent the true reason for making his

SELF-REPORTING MEMORIES OF EVENTS

What about people's memories for events in their own lifetimes? People can usually report accurately on what happened a few days ago, but what about more distant memories?

Memories for significant life experiences can be quite accurate. For example, studies have tested ways to validate adult reports of adverse childhood experiences (ACEs). One way is to locate cases of childhood abuse that were officially documented by courts or hospitals. When researchers have given ACE-related questions to court-documented victims of abuse (without revealing why they were asking), most of them did report a history of abuse, as we would expect. Other studies have shown that people accurately recalled their own abuse, even if they were not accurate about details of specific incidents (see Hardt & Rutter, 2004, for a review). The conclusion here is that people's accounts of adverse events—and many other events—should be trusted (Becker-Blease & Freyd, 2006).

In some cases, people's certainty about their memories might not match their accuracy. For example, some people failed to recall abuse that had been documented (so-called false negatives; Hardt & Rutter, 2004; Williams, 1994). They were certain they had never been victimized, despite records from their childhood that confirmed abuse.

Another example comes from people's vivid memories of exactly where they were when they heard the news that two planes had crashed into New York's World Trade Center on September 11, 2001. Cognitive psychologists have checked the accuracy of such "flashbulb memories." To conduct such a study, researchers administer a short questionnaire to people on the day after a dramatic event, asking them to recall where they were, with whom, and so forth. A few weeks or years later, the same people answer the same questions as before and also rate how vivid their memories are and how confident they are in them. People's flashbulb memories remain vivid over time (Talarico & Rubin, 2003), even as they decline in accuracy. For example, years later, about 73% of people recalling their memories of the 9/11 attacks remembered seeing the first plane hit the World Trade Center on TV, when in fact no such footage was available at that time (Pezdek, 2003).

The important finding from these studies is that vividness and confidence are unrelated to how accurate the memories actually are. Years later, people who are extremely confident in their memories are about as likely to be wrong as people who report their memories with little or no confidence (Neisser & Harsch, 1992). People's feelings of confidence in their memories do not, by themselves, inform us about their accuracy.

RATING CONSUMER PRODUCTS

What about the special case of online product ratings? Online ratings are examples of frequency claims. Are consumers able to make good judgments about products they have purchased and used? One study found little correspondence between five-star ratings on Amazon.com and the ratings of the same products by Consumer Reports, an independent product rating firm (De Langhe et al., 2016). The researchers found that consumers' ratings were, instead, correlated with the cost of the product and the prestige of its brand. Studies like these suggest that people may not always be able to accurately report on the quality of products they buy (**Figure 6.6**).

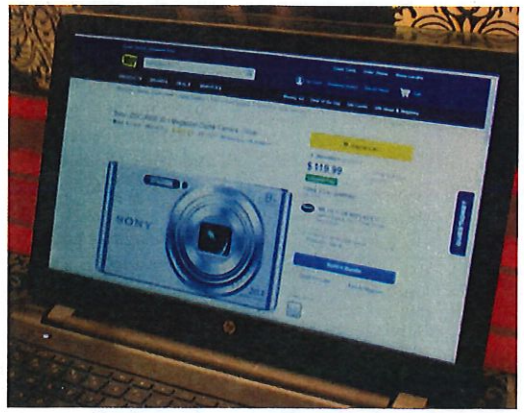


FIGURE 6.6
Do consumer ratings match expert ratings?
This camera's online reviews were positive, but Consumer Reports (an independent rating firm) ranked it second to last. While consumers may be able to report their subjective experience with a product, their ratings might not accurately predict product quality.

CHECK YOUR UNDERSTANDING

1. What are three potential problems related to the wording of survey questions? Can they be avoided?
2. Name at least two ways to ensure that survey questions are answered accurately.
3. For which topics, and in what situations, are people most likely to answer survey questions accurately ?

1. See pp. 156–159. 2. See pp. 160–163. 3. See pp. 159–160.

CONSTRUCT VALIDITY OF BEHAVIORAL OBSERVATIONS

Survey and poll results are among the most common types of data used to support a frequency claim—the kind you read most often in newspapers or on websites. Researchers also study people simply by watching them in action. When a researcher watches people or animals and systematically records how they behave or what they are doing, it is called **observational research**. Because people cannot always report on their behavior or past events accurately, some scientists believe observing behavior is better than collecting self-reports through surveys. Given the potential effects of question order, response sets, socially desirable responding, and other problems, many psychologists trust behavioral data more than survey data, at least for some variables.

Observational research can be the basis for frequency claims. Researchers might record how much people eat in fast-food restaurants or observe whether drivers stop for a pedestrian in a crosswalk. They might test the balance of athletes who have been hit on the head during practice or watch families as they eat dinner. Observational research is not just for frequency claims: Observations can also be used to operationalize variables in association claims and causal claims. Regardless of the type of claim, it is important that observational measures have good construct validity.

Some Claims Based on Observational Data

Self-report questions can be excellent measures of what people *think* they are doing and of what they *think* is influencing their behavior. But if you want to know what people are *really* doing or what *really* influences their behavior, you should probably watch them. Here are three examples of how observational methods have been used to answer research questions in psychology.

OBSERVING HOW MUCH PEOPLE TALK

Matthias Mehl and his colleagues kept track of what people say in everyday contexts (Mehl et al., 2007). The researchers recruited several samples of students and asked them to wear an electronically activated recorder (EAR) for 2–10 days (depending on the sample). This device contains a small, clip-on microphone and a digital sound recorder similar to an iPod (**Figure 6.7A**). At 12.5-minute intervals throughout the day, the EAR records 30 seconds of ambient sound. Later, research assistants transcribe everything the person says during the recorded time periods. The published data demonstrate that, on average, women spoke 16,215 words per day, while men spoke 15,669 (**Figure 6.7B**). This difference is tiny (about 3% more), so despite stereotypes of women being the chattier gender, women and men showed the same level of speaking, at least in this sample.

OBSERVING WHERE BABIES AND CAREGIVERS LOOK

A team of researchers investigated babies and their parents during play to determine how much time they spent looking at each other's faces versus looking at the toys. Individual parents and their 12-month-old babies participated in pairs. Each was fitted with a special device that recorded their eye movements while

A

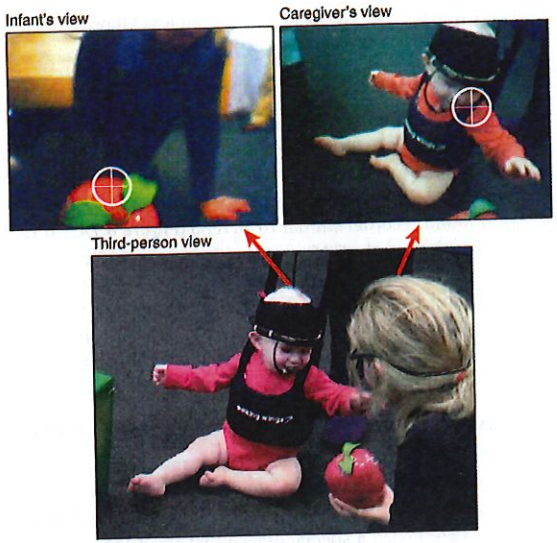
B

STRAIGHT FROM THE SOURCE

Table 1. Estimated number of words spoken per day for female and male study participants across six samples. $N = 396$. Year refers to the year when the data collection started; duration refers to the approximate number of days participants wore the EAR; the weighted average weighs the respective sample group mean by the sample size of the group.

Sample	Year	Location	Duration	Age range (years)	Sample size (*N*)		Estimated average number (SD) of words spoken per day	
					Women	Men	Women	Men
1	2004	USA	7 days	18–29	56	56	18,443 (7460)	16,576 (7871)
2	2003	USA	4 days	17–23	42	37	14,297 (6441)	14,060 (9065)
3	2003	Mexico	4 days	17–25	31	20	14,704 (6215)	15,022 (7864)
4	2001	USA	2 days	17–22	47	49	16,177 (7520)	16,569 (9108)
5	2001	USA	10 days	18–26	7	4	15,761 (8985)	24,051 (10,211)
6	1998	USA	4 days	17–23	27	20	16,496 (7914)	12,867 (8343)
					Weighted average		16,215 (7301)	15,669 (8633)

FIGURE 6.7
Observational research on daily spoken words.
(**A**) Study participants wore a small recording device to measure how many words they spoke per day. (**B**) This table shows the study's results as they were reported in the original empirical journal article. (Source: Mehl et al., 2007, Table 1.)



i.8
e trackers in observational

inel: A baby and her mother wear neras that detect exactly where are directed. **Upper left:** This ows the infant's viewpoint. The indicate the baby is looking at the **Upper right:** This camera shows the iewpoint. The crosshairs indicate t is looking at her baby's face. ranchak et al., 2016.)

they played freely in a laboratory playroom (Franchak et al., 2016; **Figure 6.8**). Later, the researchers looked frame-by-frame at the resulting video and recorded where each person was looking. One key result was that babies spent much more time looking at toys than at their parent, but that caregivers looked equally at the toys and their baby.

OBSERVING FAMILIES IN THE EVENING

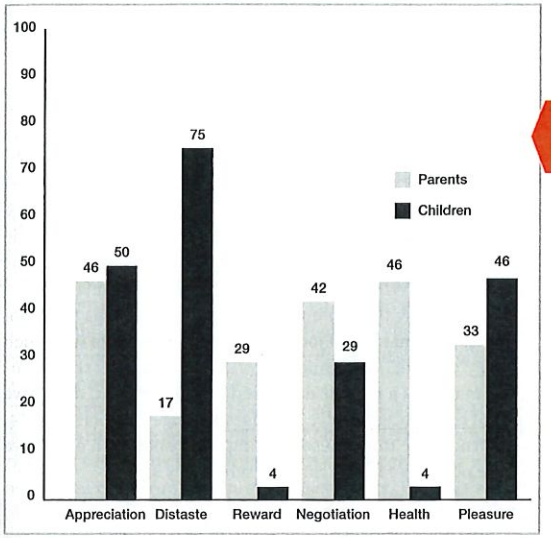
A third example comes from a study of families in which both parents work (Campos et al., 2013). The researchers had camera crews follow both parents from a sample of 30 dual-earner families from the time they got home from work until 8:00 PM. Later, teams of assistants coded a variety of behaviors from the resulting videotapes. The researchers studied two aspects of family life: the emotional tone of the parents and the topics of conversation during dinner.

To code emotional tone, they watched the videos, rating each parent on a 7-point scale. The rating scale went from 1 (cold/hostile) to 4 (neutral) to 7 (warm/happy). The results showed that emotional tone in the families was very slightly positive in the evening hours (around 4.2 on the 7-point scale). In addition, they found that kids and parents differed in what they discussed at dinnertime. The kids were more likely to express distaste at the food, while the parents talked about how healthy it was (**Figure 6.9**). In addition to these frequency estimates, the researchers also studied associations. For example, they found that mothers' (but not fathers') emotional tone was more negative when children complained about the food at dinner.

A

Dinnertime talk. Coders documented whether each family member present engaged in each of the following six types of food-related talk: (a) expressions of appreciation, (b) expressions of distaste, (c) reference to health, (d) reference to pleasure, (e) reference to food as a reward, and (f) negotiation over the terms of food rewards or penalties.

B



STRAIGHT FROM THE SOURCE

FIGURE 6.9
Coding dinnertime topics.
(A) The coders assigned each piece of dinnertime conversation to one of six categories. **(B)** The results showed that children were most likely to express distaste at the food their parents had prepared. (Source: Adapted from Campos et al., 2013.)

OBSERVATIONS CAN BE BETTER THAN SELF-REPORTS

The previous examples illustrate a variety of ways researchers have conducted observational studies. Let's reflect on the benefits of behavioral observation in these cases. What might have happened if the researchers had asked the participants to self-report? The college students certainly would not have been able to estimate how many words they spoke each day. Parents might report that they look more at their adorable babies than at toys, but the babies could not have reported at all! And while parents could report on how they were feeling in the evening and at dinner, they might not have been able to describe how emotionally warm their expressions appeared to others—the part that matters to their partners and children. Observations can sometimes tell a more accurate story than self-reporting (Vazire & Carlson, 2011).

Making Reliable and Valid Observations

Observational research is a way to operationalize a conceptual variable, so when interrogating a study we need to ask about the construct validity of any observational measure. We ask: What is the variable of interest, and did the observations accurately measure that variable? Although observational research may seem straightforward, researchers must work diligently to be sure their observations are reliable and valid.

The construct validity of observations can be threatened by three problems: observer bias, observer effects, and reactivity. Observations have good construct validity to the extent that they can avoid these three problems.

OBSERVER BIAS: WHEN OBSERVERS SEE WHAT THEY EXPECT TO SEE

Observer bias occurs when observers' expectations influence their interpretation of the participants' behaviors or the outcome of the study. Instead of rating behaviors objectively, observers rate behaviors according to their own expectations or hypotheses. In one study, psychoanalytic therapists were shown a videotape of a 26-year-old man talking to a professor about his feelings and work experiences (Langer & Abelson, 1974). Some of the therapists were told the young man was a patient, while others were told he was a job applicant. After seeing the videotape, the clinicians were asked for their observations. What kind of person was this young man?

Although all the therapists saw the same videotape, their reactions were not the same. Those who thought the man was a job applicant described him with such terms such as "attractive," "candid," and "innovative." Those who saw the videotape thinking the young man was a patient described him as a "tight, defensive person" and "frightened of his own aggressive impulses" (Langer & Abelson, 1974, p. 8). Since everyone saw the same tape, these striking differences can only have reflected the biases of the observers in interpreting what they saw.

OBSERVER EFFECTS: WHEN PARTICIPANTS CONFIRM OBSERVER EXPECTATIONS

It is problematic when observer biases affect researchers' own interpretations of what they see. It is even worse when the observers inadvertently change the behavior of those they are observing, such that participant behavior changes to match observer expectations. Known as **observer effects**, or *expectancy effects*, this phenomenon can occur even in seemingly objective observations.

Bright and Dull Rats. In a classic study of observer effects, researchers Rosenthal and Fode (1963) gave each student in an advanced psychology course five rats to test as part of a final lab experience in the course. Each student timed how long it took for their rats to learn a simple maze, every day for several days. Although each student received a randomly selected group of rats, the researchers told half of them that their rats were bred to be "maze-bright" and the other half that their rats were bred to be "maze-dull."

Even though all the rats were genetically similar, those that were believed to be maze-bright completed the maze a little faster each day and with fewer mistakes. In contrast, the rats believed to be maze-dull did not improve their performance over the testing days. Each group of rats behaved in ways that matched their observers' expectations. This study showed that observers not only see what they expect to see; sometimes they even cause the behavior of those they are observing to conform to their expectations.

Das lesende und rechnende Pferd, mit seinem Lehrer HERRN von OSTEN (Berlin)

FIGURE 6.10
William von Osten and Clever Hans.
The horse Clever Hans could detect nonverbal gestures from anybody—not just his owner—so his behavior even convinced a special commission of experts in 1904.

Clever Hans. A horse nicknamed Clever Hans provides another classic example of how observers' subtle behavior changed a subject's behavior—and how scientifically minded observers corrected the problem (Heinzen et al., 2015). More than 100 years ago, a retired schoolteacher named William von Osten tutored his horse, Hans, in mathematics. If he asked Hans to add 3 and 2, for example, the horse would tap his hoof five times and then stop. After 4 years of daily training, Clever Hans could perform math at least as well as an average fifth-grader, identify colors, and read German words (**Figure 6.10**). Von Osten allowed many scientists to test his horse's abilities, and all were satisfied that von Osten was not giving Hans cues on the sly because the horse apparently could do arithmetic even when his owner was not present.

Just when other scientists had concluded Clever Hans was truly capable of doing math, an experimental psychologist, Oskar Pfungst, came up with a more rigorous set of checks (Pfungst, 1911). Suspecting the animal was sensing subtle nonverbal cues from his human questioners, Pfungst showed the horse a series of cards printed with numbers. He alternated situations in which the questioner could or could not see each card. As Pfungst suspected, Hans was correct only when his questioner saw the card.

As it turned out, the horse *was* extremely clever—but not at math. He was smart at detecting the subtle head movements of the questioner. Pfungst noticed that a questioner would lean over to watch Hans tap his foot, raising his head a bit at the last correct tap. Clever Hans had learned this slight move was the cue to stop tapping (Heinzen et al., 2015).

PREVENTING OBSERVER BIAS AND OBSERVER EFFECTS

Researchers must ensure the construct validity of observational measures by taking steps to avoid observer bias and observer effects. First and foremost, careful researchers train their observers well. They develop clear rating instructions, often called *codebooks*, so the observers can make reliable judgments with minimal bias. Codebooks are precise statements of how the variables are operationalized, and the more precise and clear the codebook statements are, the more valid the

Emotional tone. To measure emotional tone, coders rated the extent to which the behavior of each parent was marked by verbal and nonverbal markers of coldness/hostility or warmth/happiness on a Likert scale (1 = *cold/hostile*; 4 = *neutral*; 7 = *warm/happy*). Cold/hostile emotional tone was defined as short communication, flat or angry affect, and no evidence of positive affect. Neutral tone was defined as a task oriented, practical tone that was neither cold/hostile nor warm/happy. Warm/happy emotional tone was defined as warm voice tones, smiles, laughter, and head nods with no evidence of negative affect. Coders independently rated a parent's emotional tone when they appeared in the video (a) alone, (b) with their partner (if present), or (c) with their 7- to 12-year-old child (if present). The latter rating was restricted to the 7- to 12-year-old child that all families were required to have to standardize interaction that might otherwise vary with stage of child development. Thus, up to three emotional tone variables could be rated for each parent in each 30-s video slice. Interrater reliabilities for emotional tone alone (ICC = .92), with partner (ICC = .91), and with child (ICC = .95) were high.

STRAIGHT FROM THE SOURCE

Interrater reliability of this observation.

FIGURE 6.11
Coding emotional tone.
Here is how the researchers reported the way they coded emotional tone in the Method section of their article. (Source: Adapted from Campos et al., 2013.)

operationalizations will be. **Figure 6.11** shows how emotional tone was coded in the family observation study.

Researchers can assess the construct validity of a coded measure by using multiple observers. Doing so allows the researchers to assess the interrater reliability of their measures. The excerpt shown in Figure 6.11 shows how researchers discuss the interrater reliability of the emotional tone ratings. The abbreviation ICC is a correlation that quantifies degree of agreement. The closer the correlation is to 1.0, the more the observers agreed with one another. The coders in this case showed acceptable interrater reliability.

» more on interrater ility, see Chapter 5, pp. 125–132.

Using multiple observers does not eliminate anyone's biases, of course, but if two observers of the same event agree on what happened, the researchers can be more confident. If there is disagreement, the researchers may need to train their observers better, develop a clearer coding system for rating the behaviors, or both.

Even when an operationalization has good interrater reliability, it still might not be valid. When two observers agree with each other, they might share the same biases, so their common observations are not necessarily valid. Think about the therapists in the Langer and Abelson (1974) study. Those who were told the man in the videotape was a patient might have showed interrater reliability in their descriptions of how defensive or frightened he appeared. But because they shared similar biases, their reliable ratings were not valid descriptions of the man's behavior. Therefore, interrater reliability is only half the story; researchers should employ methods that minimize observer bias and observer effects.

Masked Research Design. The Rosenthal and Fode (1963) study and the Clever Hans effect both demonstrate that observers can give unintentional cues that influence the behavior of their subjects. A common way to prevent observer bias and observer effects is to use a **masked design**, or *blind design*, in which the observers are unaware of the purpose of the study and the conditions to which participants have been assigned.

If Rosenthal and Fode's students had not known which rats were expected to be bright and dull, the students would not have evoked different behavior in their charges. Similarly, when Clever Hans' observers did not know the right answer to the questions they were asking, the horse acted differently and seemed much less intelligent. These examples make it clear that coders and observers should not be aware of a study's hypotheses, or steps should be taken to mask the conditions they are observing.

REACTIVITY: WHEN PARTICIPANTS REACT TO BEING WATCHED

Sometimes the mere presence of an outsider is enough to change the behavior of those being observed. Suppose you're visiting a first-grade classroom to observe the children. You walk quietly to the back of the room and sit down to watch what the children do. What will you see? A roomful of little heads swiveled around looking at you! Do first graders usually spend most of their time staring at the back of the room? Of course not. What you are witnessing is an example of reactivity.

Reactivity is a change in behavior when study participants know another person is watching. They might react by being on their best behavior—or in some cases, their worst—rather than displaying their typical behavior. Reactivity occurs not only with human participants but also with animal subjects. If people and animals can change their behavior just because they are being watched, what should a careful researcher do?

Solution 1: Blend In. One way to avoid observer effects is to make **unobtrusive observations**—that is, make yourself less noticeable. A developmental psychologist doing research might sit behind a one-way mirror, like the one shown in **Figure 6.12**, in order to observe how children interact in a classroom without letting them know. In a public setting, a researcher might act like a casual onlooker—another face in the crowd—to watch how other people behave.

Solution 2: Wait It Out. Another solution is to wait out the situation. A researcher who plans to observe at a school might let the children get used to his or her presence until they forget they're being watched. The anthropologist Jane Goodall, in her studies of chimpanzees in the wild, used a similar tactic. When she began introducing herself to the chimps in the

FIGURE 6.12
Unobtrusive observations.
This one-way mirror lets researchers unobtrusively record the behaviors of children in a preschool classroom.

Gombe National Park in Africa, they fled or stopped whatever else they were doing to focus on her. After several months, the chimps got used to having her around and were no longer afraid to go about their usual activities in her presence. Similarly, participants in the Mehl EAR study reported that after a couple of days of wearing the device, they did not find it to be invasive (Mehl & Pennebaker, 2003).

Solution 3: Measure the Behavior's Results. Another way to avoid reactivity is to use unobtrusive data. Instead of observing behavior directly, researchers measure the traces a particular behavior leaves behind. For example, in a museum, wear-and-tear on the flooring can signal which areas of the museum are the most popular, and the height of smudges on the windows can indicate the age of visitors. The number of empty liquor bottles in residential garbage cans indicates how much alcohol is being consumed in a community (Webb et al., 1966). Using these indirect methods, researchers can measure behavior without doing any direct participant observation.

OBSERVING PEOPLE ETHICALLY

Is it ethical for researchers to observe the behaviors of others? It depends. Most psychologists believe it is ethical to watch people in museums and classrooms, at sports events, or even at the sinks of public bathrooms because in those settings people can reasonably expect their activities to be public, not private. Of course, when psychologists report the results of such observational studies, they do not specifically identify any of the people who were observed.

More secretive methods, such as one-way mirrors and covert video recording, are also considered ethical in some conditions. In most cases, psychologists doing research must obtain permission in advance to watch or to record people's private behavior. If hidden video recording is used, the researcher must explain the procedure at the conclusion of the study. If people object to having been recorded, the researcher must erase the file without watching it.

Certain ethical decisions may be influenced by the policies of a university where a study is conducted. As discussed in Chapter 4, institutional review boards (IRBs) assess each study to decide whether it can be conducted ethically.

CHECK YOUR UNDERSTANDING

1. Sketch a concept map of observer bias, observer effects, and reactivity, and indicate the approaches researchers can take to minimize each problem.
2. Explain why each of these three problems can threaten construct validity, using this sentence structure for each issue:

   If an observational study suffers from _____, then the researcher might be measuring _____ instead of _____.

1. See pp. 170–174. 2. See pp. 170 and 172.

CHAPTER REVIEW

It's time to complete your study experience! Go to INQUIZITIVE to practice actively with this chapter's concepts and get personalized feedback along the way.

Summary

Surveys, polls, and observational methods are used to support frequency claims, but they also measure variables for association and causal claims. When interrogating a claim based on data from a survey or an observational study, we ask about the construct validity of the measurement.

CONSTRUCT VALIDITY OF SURVEYS AND POLLS

- Survey question formats include open-ended, forced-choice, Likert scale, and semantic differential.
- Sometimes the way a survey question is worded can lead people to be more likely or less likely to agree with it.
- Double-barreled and negatively worded questions are difficult to answer in a valid way.
- People sometimes answer survey questions with an acquiescent or fence-sitting response tendency or in a way that makes them look good. To avoid some of these problems, researchers can add items to a survey or change the way questions are written.
- Surveys are efficient and accurate ways to assess people's subjective feelings and opinions; they may be less appropriate for assessing people's actual behavior, motivations, or certain memories.

CONSTRUCT VALIDITY OF BEHAVIORAL OBSERVATIONS

- Observational studies record people's true behavior, rather than what people say about their behavior.
- Well-trained coders and clear codebooks help ensure that observations will be reliable and not influenced by observer expectations.
- Some observational studies are susceptible to reactivity. Masked designs and unobtrusive observations make it more likely that observers will not make biased ratings, and that participants will not change their behavior in reaction to being observed.
- Local IRB guidelines may vary, but in general, it is considered ethical to conduct observational research in public settings where people expect to be seen by others.

Key Terms

survey, p. 154
poll, p. 154
open-ended question, p. 154
forced-choice question, p. 154
Likert scale, p. 155
semantic differential format, p. 155
leading question, p. 156
double-barreled question, p. 157
negatively worded question, p. 158
response set, p. 161
acquiescence, p. 161
fence sitting, p. 162
socially desirable responding, p. 162
faking good, p. 162
faking bad, p. 163
observational research, p. 166
observer bias, p. 170
observer effect, p. 170
masked design, p. 173
reactivity, p. 173
unobtrusive observation, p. 173

see samples of chapter concepts in the popular media,
t www.everydayresearchmethods.com and click the box for Chapter 6.

Questions

wing item appears on a survey: "Was your
e purchased within the last two years and
downloaded the most recent updates?"
he biggest problem with this wording?

leading question.

lves negative wording.

double-barreled question.

ot on a Likert scale.

ople are using an acquiescent response set

g to give the responses they think the
rcher wants to hear.

presenting their views to appear more
lly acceptable.

g the same, neutral answer to each question.

ing to agree with every item, no matter what
's.

1 of the following situations do people most
ely answer survey questions?

n they are describing the reasons for their
behavior.

n they are describing what happened to
1, especially after important events.

n they are describing their subjective experi-
e; how they personally feel about something.

ple almost never answer survey questions
urately.

4. Which of the following makes it more likely that behavioral observations will have good interrater reliability?
   a. A masked study design
   b. A clear codebook
   c. Using naive, untrained coders
   d. Open-ended responses

5. Which one of the following is a means of controlling for observer bias?
   a. Using unobtrusive observations.
   b. Waiting for the participants to become used to the observer.
   c. Making sure the observer does not know the study's hypotheses.
   d. Measuring physical traces of behavior rather than observing behavior directly.

6. Which of the following is a way of preventing reactivity?
   a. Waiting for the participants to become used to the observer.
   b. Making sure the observers do not know the study's hypotheses.
   c. Making sure the observer uses a clear codebook.
   d. Ensuring the observers have good interrater reliability.

ing Actively

er the various survey question formats: open-
forced-choice, Likert scale, and semantic
ntial. For each of the following research topics,
question in each format, keeping in mind
of the pitfalls in question writing. Which of the
ons you wrote would have the best construct
y, and why?

tudy that measures attitudes about women
ving in combat roles in the military.

   b. A customer service survey asking people about their satisfaction with their most recent shopping experience.
   c. A poll that asks people which political party they have supported in the past.

2. As part of their Well-Being Index, the Gallup organization asks a daily sample of Americans, "In the last seven days, on how many days did you exercise for 30 or more minutes?" If people say they have exercised three or more days, Gallup classifies them as "frequent exercisers." Gallup finds that between about 47% (in the winter months) and 55% (in the summer) report being frequent exercisers (Gallup, n.d.). What kind of question is this: forced-choice, Likert scale, semantic differential, or some other format? Does the item appear to be leading, negatively worded, or double-barreled? Do you think it leads to accurate responses?

3. Plan an observational study to see which kind of drivers are more likely to stop for a pedestrian in a crosswalk: male or female drivers. Think about how to maximize your construct validity. Will observers be biased about what they record? How might they influence the people they're watching, if at all? Where should they stand to observe driver behavior? How will you evaluate the interrater reliability of your observers? Write a two- to three-sentence operational definition of what it means to "stop for a pedestrian in a crosswalk." The definition should be clear enough that if you asked two friends to use it to code "stopping for pedestrian" behavior, it would have good reliability and validity.

4. To study the kinds of faces babies usually see, researchers asked parents to place tiny video cameras on their 1-month-old and 3-month-old infants during their waking hours (Sugden et al., 2013). Coders viewed the resulting video footage frame by frame, categorizing the gender, race, and age of the faces each baby saw. The results revealed that babies are exposed to faces 25% of the time, and 96% of the time they were exposed to faces that were the same race as themselves. What questions might you ask to decide whether the observational measures in this study were susceptible to observer bias, observer effects, or reactivity?