CHAPTER SEVEN

QUESTIONS AND ANSWERS IN SURVEYS

questionnaire

As we pointed out in Chapter 5, surveys use a variety of methods to collect information about their respondents. Perhaps the most common is the use of the questionnaire, a standardized set of questions administered to the respondents in a survey. The questions are typically administered in a fixed order and often with fixed answer options. Usually, interviewers administer a questionnaire to the respondents, but many surveys have the respondents complete the questionnaires themselves. Among our six example surveys, three (the NCVS, the SOC, and the BRFSS) rely almost exclusively on interviewer administration of the questions, and the other three (the NSDUH, the NAEP, and the CES) mix interviewer administration with self-completion of the questions by the respondents. Over the last 30 years or so, survey questionnaires have increasingly taken electronic form, with computer programs displaying the questions either to the interviewer or directly to the respondent. But whether an interviewer is involved or not, and whether the questionnaire is a paper document or a computer program, most surveys rely heavily on respondents to interpret a preestablished set of questions and to supply the information these questions seek.

7.1 Alternatives Methods of Survey Measurement

Surveys do not always force respondents to construct answers during the interview. For example, many surveys collect information from businesses or other establishments, and such surveys often draw information from company records. For these, the questionnaire may be more like a data recording form than like a script for the interview, and the interviewer may interact with the records rather than with a respondent. (The CES survey sometimes comes close to this model.) Similarly, education surveys may supplement data collected via questionnaires with data from student transcripts; and health surveys may extract information from medical records instead of relying completely on respondents' reports about their medical treatment or diagnosis. Even when the records are the primary source of the information, the respondent may still play a key role in gaining access to the records and in helping the data collection staff extract the necessary information from them. Some surveys collect the data during interviews but ask respondents to assemble the relevant records ahead of time to help them supply accurate answers. For instance, the National Medical Expenditure Survey and its successor, the Medical Expenditure Panel Survey, encouraged the respondents to keep doctor bills and other records handy to help answer questions about doctor

visits and medical costs during the interviews. Sometimes, records might be extremely helpful to the respondents as they answer the survey questions, but the necessary records just do not exist. For example, few households keep detailed records of their everyday expenses; if they did, they would be in a much better position to provide the information sought by the U.S. Bureau of Labor Statistics in its Consumer Expenditure Survey (CES), which tracks household spending. These and other surveys attempt to persuade the respondents to create contemporaneous record-like data by keeping diaries of the relevant events. Like surveys that rely on existing records, diary surveys shift the burden from the respondents' memories to their record keeping.

Another type of measurement used in surveys involves standardized psychological assessments. Many educational surveys attempt to relate educational outcomes to characteristics of the schools, teachers, or parents of the students; cognitive tests are used in such studies to provide comparable measurements across sample students. One of our example surveys, the National Assessment of Educational Progress (NAEP), leans heavily on standardized tests of academic achievement for its data.

This chapter focuses on the issues raised by survey questionnaires. Virtually all surveys use questionnaires, and even when surveys do not use questionnaires, they still rely on standardized data collection instruments, such as record abstraction forms or diaries. Many of the principles involved in creating and testing questionnaires apply to these other types of standardized instruments as well.

7.2 Cognitive Processes in Answering Questions

Almost all surveys involve respondents answering questions put to them by an interviewer or a self-administered questionnaire. Several researchers have attempted to spell out the mental processes set in motion by survey questions. Most of the resulting models of the response process include four groups of processes: "comprehension" (in which respondents interpret the questions), "retrieval" (in which they recall the information needed to answer them), "judgment" (in which they combine or summarize the information they recall), and "reporting" (in which they formulate their response and put it in the required format). See Figure 7.1.

encoding

In some cases, it is also important to take into account the cognitive processes that take place before the interview, when the respondent first experiences the

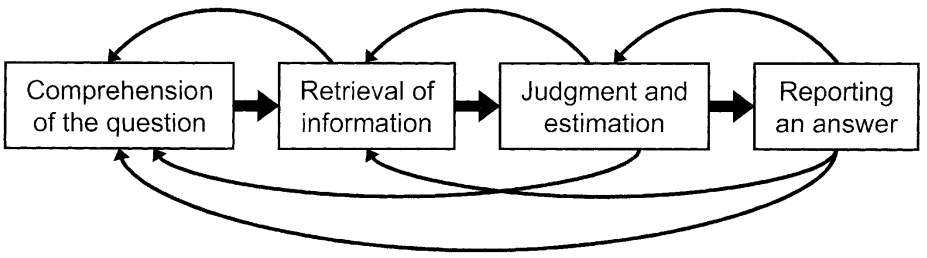

Figure 7.1 A simple model of the survey response process.

events in question. "Encoding" is the process of forming memories from experiences. Generally, survey designers have little impact on these processes, but there is good evidence that survey questions can be improved if they take into account how respondents have encoded the information the survey questions seek to tap.

Finally, with self-administered questions, respondents also have to figure out how to work their way through the questionnaire, determining which item comes next and digesting any other instructions about how to complete the questions. We treat this process of interpreting navigational cues and instructions as part of the comprehension process. (For early examples of models of the survey response process, see Cannell, Miller, and Oksenberg, 1981; and Tourangeau, 1984. For more up-to-date treatments, see Sudman, Bradburn, and Schwarz, 1996; and Tourangeau, Rips, and Rasinski, 2000.)

It is easy to give a misleading impression of the complexity and rigor of the cognitive processes involved in answering survey questions. Much of the evidence reviewed by Krosnick (1999), Tourangeau, Rips, and Rasinski (2000), and others who have examined the survey response process in detail indicates that respondents often skip some of the processes listed in Figure 7.1 and that they carry out others sloppily. The questions in surveys are often difficult and place heavy demands on memory or require complicated judgments. Consider these items drawn from our example surveys below. We follow the typographical conventions of the surveys themselves. For example, the blank slot in the NCVS item (used until 2000) is filled in at the time of the interview with an exact date.

> On average, during the last 6 months, that is, since _____, how often have YOU gone shopping? For example, at drug, clothing, grocery, hardware and convenience stores. [NCVS]

> Now turning to business conditions in the country as a whole, do you think that during the next 12 months we'll have good times financially, or bad times, or what? [SOC]

Imagine how difficult it would be to come up with an exact and accurate answer to the question on shopping posed on the NCVS. Fortunately, the NCVS is looking only for broad categories, like "once a week," rather than a specific number. Even so, the memory challenges this item presents are likely to be daunting and most respondents probably provide only rough estimates for their answers. Similarly, most of us are not really in a position to come up with thoughtful forecasts about "business conditions" (whatever that means!) during the next year. There is no reason to think that the typical respondents in surveys have either the time or the inclination to work hard to answer such questions, and there is ample evidence that they take lots of shortcuts to simplify their task.

We also do not mean to give the impression that respondents necessarily carry out the relevant cognitive processes in a fixed sequence of steps beginning with comprehension of the question and ending with reporting an answer. Some degree of backtracking and overlap between these processes is probably the rule rather than the exception. In addition, although most of us have some experience with surveys and recognize the special conventions that surveys employ (such as five-point response scales), we also have a lifetime of experience in dealing with a wide range of questions in everyday life. Inevitably, the habits and the strategies

we have developed in other situations carry over in our attempts to make sense of survey questions and to satisfy the special demands that survey questions make.

7.2.1 Comprehension

comprehension

"Comprehension" includes such processes as attending to the question and accompanying instructions, assigning a meaning to the surface form of the question, and inferring the question's point (i.e., identifying the information sought). For concreteness, let us focus our discussion of how people answer survey questions on a specific item from the NSDUH:

> Now think about the past 12 months, from [DATE] through today. We want to know how many days you've used any prescription tranquilizer that was not prescribed to you or that you took only for the experience or feeling it caused during the past 12 months. [NSDUH]

It is natural to suppose that the respondent's first task is to interpret this question. (Although this item is not in the interrogative form, it nonetheless conveys a request for information and will generally be treated by respondents in the same way a grammatical question would be.) Like most survey items, this one not only requests information about a particular topic, in this case, the illicit use of prescription tranquillizers, it also requests that the information be provided in a specific form, ideally the number (between one and 365) corresponding to the number of days the respondent has used a certain type of prescription drugs in a certain way. (Respondents who reported that they had not used prescription tranquilizers illicitly during the past year do not get this item, so the answer should be at least one day.) The next part of the question (which is not shown) goes further in specifying the intended format of the answer, offering the respondents the choice of reporting the average days per week, the average days per month, or the total days over the whole year. Thus, one key component of interpreting a question (some would argue *the* key component) is determining the set of permissible answers; surveys often offer the respondents help in this task by providing them with a fixed set of answer categories or other clear guidelines about the form the responses should take.

Interpretation is likely to entail such component processes as parsing the question (identifying its components and their relations to each other), assigning meanings to the key substantive elements (terms like "used" and "prescription tranquilizers"), inferring the purpose behind the question, and determining the boundaries and potential overlap among the permissible answers. In carrying out these interpretive tasks, respondents may go beyond the information provided by the question itself, looking to the preceding items in the questionnaire or to the additional cues available from the interviewer and the setting. And, although it is obvious that the NSDUH item has been carefully framed to avoid confusion, it is easy to see how respondents might still have trouble with the exact drugs that fall under the heading of "prescription tranquilizers" or how they might differ in implementing the definitional requirements for using such drugs illicitly. Even everyday terms like "you" in the NCVS example above can cause problems—does "you" mean you personally or you and other members of your household?

(The use of all capital letters in the NCVS is probably intended to convey to the interviewer that the item covers only the respondent and not the other persons in the household.)

7.2.2 Retrieval

retrieval

"Retrieval" is the process of recalling information relevant to answering the question from long-term memory. Long-term memory is the memory system that stores autobiographical memories, as well as general knowledge. It has a vast capacity and can store information for a lifetime.

To determine which of the possible answers to the NSDUH item is actually true, the respondent will typically draw on his or her memory for the type of event in question. The study of memory for autobiographical events is not much older than the study of how people answer survey questions, and there is not complete agreement on how memory for autobiographical information is organized (see Barsalou, 1988, and Conway, 1996, for two attempts to spell out the relevant structures). Nonetheless, there is some consensus about how people retrieve information about the events they have experienced. Survey respondents might begin with a set of cues that includes an abstract of the question itself (e.g., "Hmm, I've used tranquilizers without a prescription a number of times over the past year or so"). A retrieval cue is a prompt to memory. It provides some clue that helps trigger the recall of information from long-term memory. This initial probe to memory may call up useful information from long-term memory—in the best case, an exact answer to the question itself. In many situations, however, memory will not deliver the exact answer but will provide relevant information, in the form of further cues that help lead to an answer (e.g., "It seems like I took some pills that Lance had one time, even though I didn't really need them"). Inferences and other memories based on these self-generated cues (e.g., "If Lance had a prescription, it probably contained about fifty tablets or so") help constrain the answer and offer further cues to probe memory. This cycle of generating cues and retrieving information continues until the respondent finds the necessary information or gives up.

retrieval cue

Several things affect how successful retrieval is. In part, success depends on the nature of the events in question. Some things are generally harder to remember than others; when the events we are trying to remember are not very distinctive, when there are a lot of them, and when they did not make much of an impression in the first place, we are hard put to remember them. If you abused prescription tranquilizers often, over a long period of time, and in a routine way, chances are you will find it difficult to remember the exact number of times you did so or the details of the specific occasions when you used tranquilizers illicitly. On the other hand, if you have abused tranquilizers just once or twice, and you did so within the last few days, retrieval is more likely to produce the exact number and circumstances of the incidents.

Of course, another major factor determining how difficult or easy it is to remember something is how long ago it took place. Psychologists who have studied memory have known for more than 100 years that memory gets worse as the events to be remembered get older. Our example from the NSDUH uses a relatively long time frame (a year) and that may make it difficult for some respondents to remember all the relevant incidents, particularly if there are a lot of them.

Another factor likely to affect the outcome of retrieval is the number and richness of the "cues" that initiate the process. The NCVS item on shopping tries to offer respondents help in remembering by listing various kinds of stores they may have visited ("drug, clothing, grocery, hardware and convenience stores"). These examples are likely to cue different memories. The best cues are the ones that offer the most detail, provided that the specifics in the cue match the encoding of the events. If, for example, in the NSDUH question a respondent does not think of Valium as a "prescription tranquilizer," then it may not tap the right memories. Whenever the cues provided by the question do not match the information actually stored in memory, retrieval may fail.

7.2.3 Estimation and Judgment

estimation

judgment

"Estimation" and "judgment" are the processes of combining or supplementing what the respondent has retrieved. Judgments may be based on the process of retrieval (e.g., whether it was hard or easy). In addition, judgments may fill in gaps in what is recalled, combine the products of the retrieval, or adjust for omissions in retrieval.

Though the NSDUH question seems to ask for a specific number, the follow-up instructions tacitly acknowledge that the information the respondent is able to recall may take a different form, such as a typical rate. People do not usually keep a running tally of the number of times they have experienced a given type of event, so they are generally unable to retrieve some preexisting answer to questions like the one posed in the NSDUH item or the question on shopping from the NCVS. By contrast, a company might well keep a running tally of its current employees, which is the key piece of information sought in the CES. The CES also demonstrates that in some cases retrieval may involve an external search of physical records rather than a mental search of memory. The respondent could try to construct a tally on the spot by recalling and counting individual incidents, but if the number of incidents is large, recalling them all is likely to be difficult or impossible. Instead, respondents are more likely estimate the number based on typical rates. Which strategy a particular respondent adopts—recalling a tally, constructing one by remembering the specific events, giving an estimate based on a rate, or simply taking a guess—will depend on the number of incidents, the length of the period covered by the survey, the memorability of the information about specific incidents, and the regularity of the incidents, all of which will affect what information the respondent is likely to have stored in memory (e.g., Blair and Burton, 1987; and Conrad, Brown, and Cashman, 1998).

It may seem that answers to attitudinal items, like the example from the Survey of Consumers on business conditions, would require a completely different set of response processes from those required to answer the more factual NSDUH item on illicit use of prescription tranquilizers. But a number of researchers have argued that answers to attitude items also are not generally preformed, waiting for the respondent to retrieve them from memory (see, for example, Wilson and Hodges, 1992). How many respondents keep track of their views about the likely business climate in the upcoming year, updating these views as conditions change or they receive new information? Respondents are more likely to deal with issues like the one raised in the SOC as they come up,

basing their answers on whatever considerations come to mind and seem relevant at the time the question is asked (e.g., trends in unemployment and inflation, recent news about the world markets, or how the stock market has done lately). The same sorts of judgment strategies used for answering questions about behaviors have their counterparts for questions about attitudes. For example, respondents may recall specific incidents in answering the NSDUH item about tranquilizers just as they may try to remember specific facts about the economy in deciding what business conditions might bring over the next year. Or they may base their answers on more general information, like typical rates in the case of the NSDUH item or long-term economic trends in the case of the Survey of Consumers question.

7.2.4 Reporting

reporting

"Reporting" is the process of selecting and communicating an answer. It includes mapping the answer onto the question's response options and altering the answer for consistency with prior answers, perceived acceptability, or other criteria. As we already noted, our example NSDUH item not only specifies the topic of the question, but also the format for the answer. An acceptable answer could take the form of an exact number of days or a rate (days per week or per month). There are two major types of questions based on their formats for responding. A "closed" question presents the respondents with a list of acceptable answers. "Open" questions allow respondents to provide the answers in their own terms, although, typically, the answer is nonetheless quite circumscribed. Roughly speaking, open questions are like in fill-in-the-blank questions on a test; closed questions are like multiple choice questions. Attitude questions almost always use a closed format, with the answer categories forming a scale.

How respondents choose to report their answers will depend in part on the fit between the information they retrieve (or estimate) and the constraints imposed by the question. For questions that require a numerical answer, like the NSDUH item, they may have to adjust their internal judgment to the range and distribution of the response categories given. For example, if most of the response categories provided involve low frequencies, the reports are likely to be skewed in that direction. Or when no response options are given, respondents may have to decide on how exact their answer should be and round their answers accordingly. Respondents may also give more weight to certain response options, depending on the order in which they are presented (first or last in the list of permissible answers) and the mode of presentation (visual or auditory). When the topic is a sensitive one (like drug use), respondents may shade their answer up or down or refuse to answer entirely. This response censoring is more likely to happen when an interviewer administers the questions (see Section 5.3.5).

7.2.5 Other Models of the Response Process

It is worth mentioning that the simple model of the survey response process depicted in Figure 7.1 is not the only model that researchers have proposed. Cannell, Miller, and Oksenberg (1981) proposed an earlier model distinguishing

Comments on Response Strategies

Respondents may adopt broad strategies for answering a group of survey questions. Several of these strategies—selecting the "don't know" or "no opinion" answer category or choosing the same answer for every question—can greatly reduce the amount of thought needed to complete the questions. Such strategies are examples of survey "satisficing," in which respondents do the minimum they need to do to satisfy the demands of the questions.

two main routes that respondents may follow in formulating their answers. One track that includes most of the same processes that we have described here—comprehension, retrieval, judgment, and reporting—leads to accurate or at least adequate responses. The other track is for respondents who take shortcuts to get through the interview more quickly or who have motives that override their desire to provide accurate information. Such respondents may give an answer based on relatively superficial cues available in the interview situation, cues like the interviewer's appearance or the implied direction of the question. Responses based on such cues are likely to be biased by "acquiescence" (the tendency to agree) or "social desirability" (the tendency to present oneself in a favorable light by underreporting undesirable attributes and overreporting desirable ones).

A more recent model of the survey response process shares the Cannell model's assumption of dual paths to a survey response—a high road taken by careful respondents and a low road taken by respondents who answer more superficially. This is the "satisficing" model proposed by Krosnick and Alwin (1987) (see also Krosnick, 1991). According to this model, some respondents try to "satisfice" (the low road), whereas others try to "optimize" (the high road) in answering survey questions. Satisficing respondents do not seek to understand the question completely, just well enough to provide a reasonable answer; they do not try to recall everything that is relevant, but just enough material on which to base an answer; and so on. Satisficing thus resembles the more superficial branch of Cannell's two-track model. Similarly, optimizing respondents would seem to follow the more careful branch. In his later work, Krosnick has distinguished some specific response strategies that satisficing respondents may use to get through questions quickly. For example, they may agree with all attitude items that call for agree–disagree responses, a response strategy called "acquiescence."

acquiescence

social desirability

satisficing

Like the Cannell model, Krosnick's satisficing theory makes a sharp distinction between processes that probably vary continuously. Respondents may process different questions with differing levels of care, and they may not give the same effort to each component of the response process. Just because a respondent was inattentive in listening to the question, he or she will not necessarily do a poor job at retrieval. For a variety of reasons, respondents may carry out each cognitive operation carefully or sloppily. We prefer to think of the two tracks distinguished by Cannell and Krosnick as the two ends of a continuum that varies in the depth and the quality of thought respondents give in formulating their answers.

There is more research that could be done on appropriate models of the interview. One challenge of great importance is discovering the role of the computer as an intermediary in the interaction. Is a CAPI or ACASI device more like another actor in the interaction or is it more like a static paper questionnaire? What are the independent effects of computer assistance on the interviewer and the respondent in face-to-face surveys? How can software design change respondent behavior to improve the quality of survey data?

7.3 PROBLEMS IN ANSWERING SURVEY QUESTIONS

One of the great advantages of having a model of the response process, even a relatively simple one like the one in Figure 7.1, is that it helps us to think systematically about the different things that can go wrong, producing inaccurate answers to the questions. As we note elsewhere, the goal of surveys is to reduce error and one major form of error is measurement error—discrepancies between the true answer to a question and the answer that finds its way into the final database. (Although this definition of measurement error does not apply so neatly to attitude measures, we still want the responses to attitude questions to relate systematically to the underlying attitudes they are trying to tap. As a result, we prefer attitude measures that have a stronger relation to the respondents' attitudes.)

The major assumption of the cognitive analysis of survey responding is that flaws in the cognitive operations involved in producing an answer are responsible for errors in the responses. Here, we distinguish seven problems in the response process that can give rise to errors in survey reports:

1) Failure to encode the information sought.
2) Misinterpretation of the questions.
3) Forgetting and other memory problems.
4) Flawed judgment or estimation strategies.
5) Problems in formatting an answer.
6) More or less deliberate misreporting.
7) Failure to follow instructions.

There are several book-length examinations of the response process that offer longer, more detailed lists of response problems (e.g., Sudman, Bradburn, and Schwarz, 1996; Tourangeau, Rips, and Rasinski, 2000). All of these approaches share the assumption that measurement errors can generally be traced to some problem in the response process (e.g., the respondents never had the necessary information or they forgot it, they misunderstand the questions, they make inappropriate judgments, and so on).

7.3.1 Encoding Problems

The mere fact that someone has lived through an event does not necessarily mean that he or she absorbed much information about it. Studies of eyewitness testimony suggest that eye witnesses often miss key details of the unusual and involving events they are testifying about (e.g., Wells, 1993). With the more routine experiences that are the stuff of surveys, respondents may take in even less information as they experience the events. As a result, their after-the-fact accounts may be based largely on what usually happens. A study by A. F. Smith illustrates the problem. He examined survey respondents' reports about what they ate and compared these reports to detailed food diaries the respondents kept. There was such a poor match between the survey reports and the diary entries that Smith (1991, p. 11) concluded, "dietary reports ... consist in large part of individuals' guesses

Fowler (1992) on Unclear Terms in Questions

In 1992, Fowler reported a study showing that removing unclear terms during pretesting affects answers to survey questions.

Study design: About 100 pretest interviews of a 60-item questionnaire were tape-recorded. Behavior coding documented what interviewers and respondents said for each question–answer sequence. Seven questions generated calls for clarification or inadequate answers in 15% or more of the interviews. Revisions of the questions attempted to remove ambiguous words. A second round of pretesting interviewed 150 persons. Response distributions and behavior-coding data were compared across the two pretests. For example, the first pretest included the question, "What is the average number of days each week you have butter?" The second addressed the ambiguity of the word, "butter," with the change, "The next question is just about butter. Not including margarine, what is the average number of days each week that you have butter?"

Findings: The number of calls for clarification and inadequate answers declined from pretest 1 to pretest 2. Response distributions changed; for example, on the question about having butter:

	% Never Having Butter
Pretest 1	33%
Pretest 2	55%

The authors conclude that the exclusion of margarine increased those who reported never having butter.

Limitations of the study: There was no external criterion available for the true values of answers. The results provide no way of identifying what level of behavior-coding problems demand changes to question wording. The work assumes that the same question wording should be used for all.

Impact of the study: The study demonstrated that pretesting with behavior coding can identify problem questions. It showed how changes in wording of questions can improve interaction in interviews, reflected in behavior coding, and affect the resulting survey estimates.

about what they probably ate." The problem with asking people about what they eat is that most people do not pay that much attention; the result is they cannot report about it with much accuracy.

There is a practical lesson to be drawn from this example. People cannot provide information they do not have; if people never encoded the information in the first place, then no question, no matter how cleverly framed, is going to elicit accurate responses. A key issue for pretests is making sure respondents have the information the survey seeks from them.

7.3.2 Misinterpreting the Questions

Even if the respondents know the answers to the questions, they are unlikely to report them if they misunderstand the questions. Although it is very difficult to say how often respondents misunderstand survey questions, there are several indications that it happens quite frequently.

One source of evidence is a widely cited study by Belson (1981; see also Belson, 1986), who asked respondents to report what key terms in survey questions meant to them. He found that respondents assigned a range of meanings even to seemingly straightforward terms like "you" (does this cover you personally, you and your spouse, or you and your family?) or "weekend" (is Friday a weekday or part of the weekend?). Belson also studied an item whose problems probably seem a lot more obvious in retrospect:

> Do you think that children suffer any ill effects from watching programmes with violence in them, other than ordinary Westerns?

Belson's respondents gave a range of interpretations to the term "children." "Children" has two basic meanings: young people, regardless of their relation to you, and your offspring, regardless of their age. The exact age cutoff defining "children" in the first sense varies from one situation to the next (there is one age cutoff for haircuts, another for getting into R-rated movies, still another for ordering liquor), and Belson found similar variation across survey respondents. He also found some idiosyncratic definitions (e.g., nervous children, one's own grandchildren). If "children" receives multiple interpretations, then a deliberately vague term like "ill effects" is bound to receive a wide range of readings as well. (It is also not very clear why the item makes an exception for violence in "ordinary Westerns" or what respondents make of this.)

These problems in interpreting survey questions might not result in misleading answers if survey respondents were not so reluctant to ask what specific terms mean or to admit that they simply do not understand the question. Some studies have asked respondents about fictitious issues (such as the "Public Affairs Act") and found that as many as 40% of the respondents are still willing to venture an opinion on the "issue" (Bishop, Oldendick, and Tuchfarber, 1986). In everyday life, when one person asks another person a question, the assumption is that the speaker thinks it is likely or at least reasonable that the hearer will know the answer. As a result, respondents may think that they ought to know about issues like the Public Affairs Act or that they ought to understand the terms used in survey questions. When they run into comprehension problems, they may be embarrassed to ask for clarification and try to muddle through on their own. In addition, survey interviewers may be trained to discourage such questions entirely or to offer only unenlightening responses to them (such as repeating the original question verbatim). Unfortunately, as Belson's results show, even with everyday terms, different respondents often come up with different interpretations; they are even more likely to come up with a wide range of interpretations when the questions include relatively unfamiliar or technical terms.

Tourangeau, Rips, and Rasinski (2000) distinguish seven types of comprehension problems that can crop up on surveys:

1) Grammatical ambiguity.
2) Excessive complexity.
3) Faulty presupposition.
4) Vague concepts.
5) Vague quantifiers.
6) Unfamiliar terms.
7) False inferences.

grammatical ambiguity

The first three problems have to do with the grammatical form of the question. "Grammatical ambiguity" means that the question can map onto two or more underlying representations. For example, even a question as simple as "Are you visiting firemen?" can mean two different things: Are you a group of firemen that has come to visit? or Are you going to visit some firemen? In real life, context would help sort things out, but in surveys, grammatical ambiguity can produce differing interpretations across respondents. A more common problem with survey questions is excessive complexity. Here is an example discussed by Fowler (1992):

> During the past 12 months, since January 1, 1987, how many times have you seen or talked to a doctor or assistant about your health? Do not count any time you might have seen a doctor while you were a patient in a hospital, but count all other times you actually saw or talked to a medical doctor of any kind.

excessive complexity

"Excessive complexity" means that the question has a structure that prevents the respondent from inferring its intended meaning. The main question in the example above lists several possibilities (seeing a doctor, talking to an assistant), and the instructions that follow the question add to the overall complexity. The problem with complicated questions like these is that it may be impossible for respondents to keep all the possibilities and requirements in mind; as a result, part of the meaning may end up being ignored.

faulty presupposition

"Faulty presupposition" means that the question assumes something that is not true. As a result, the question does not make sense or does not apply. For example, suppose respondents are asked whether they agree or disagree with the statement, "Family life often suffers because men concentrate too much on their work." The question presupposes that men concentrate too much on their work; respondents who do not agree with that assumption (for example, people who think that most men are lazy) cannot really provide a sensible answer to the question. All questions presuppose a certain picture of things and ask the respondent to fill in some missing piece of that picture; it is important that the listener does not reject the state of affairs depicted in the question.

vague concepts/ vague quantifiers

The next three problems involve the meaning of words or phrases in the question. As Belson has pointed out, many everyday terms are vague, and different respondents may interpret them differently. As a result, it helps if survey questions are as concrete as possible. For example, an item about children should specify the age range of interest. Note, however, that the attempt to spell out exactly what some vague term covers can lead to considerable complexity. That is the problem with the question above, in which the question tries to spell out the notion of an outpatient doctor visit. Some survey items employ vague relative terms ("Disagree somewhat" or "Very often") in response scales. Unfortunately, respondents may not agree about how often is very often, with the result that different respondents use the scale in different ways.

unfamiliar term

Another source of interpretive difficulty is that respondents may not know what a particular term means. The people who write questionnaires are often experts about a subject, and they may overestimate how familiar the respondents are likely to be with the terminology they themselves use every day. An economist who wants to ask people about their pension plans may be tempted to use terms like "401(k)" or "SRA" without defining them. Unfortunately, such terms are likely to confuse many respondents.

false inference

A number of findings suggest that respondents can also overinterpret survey questions, drawing false inferences about their intent. Consider this question, drawn from the General Social Survey:

> Are there any situations you can imagine in which you would approve of a *policeman* striking an adult male citizen?

It is fairly easy to imagine such circumstances leading to a "yes" answer (just watch any cop show!), but many respondents (roughly 30%) still answer "no." Clearly, many respondents do not interpret the question literally; instead, they

respond in terms of its perceived intent—to assess attitudes toward violence by the police. Such inferences about intent are a natural part of the interpretation process, but they can lead respondents astray. For example, several studies suggest that respondents (incorrectly) infer that an item about their overall happiness ("Taken altogether, how would you say things are these days? Would you say that you are very happy, pretty happy, or not too happy?") is supposed to exclude their marriages when that item comes right after a parallel item asking about marital happiness (e.g., Schwarz, Strack, and Mai, 1991). In everyday conversation, each time we speak we are supposed to offer something new; this expectation leads respondents to believe that the general item is about the rest of their lives, which has not been covered yet, apart from their marriage. Unfortunately, sometimes such inferences may not match what the survey designers intended.

This section underscores the fact that language matters. In many circumstances, there is a tension between explicitly defining terms in a question (in an attempt to eliminate ambiguity) and increasing the burden on the respondent to absorb the full intent of the question. More research is needed for determining what level of detail to offer all respondents, how to discern when definitional help is needed by a respondent, and how the interaction between interviewer and respondent affects the interpretation of the verbal content of questions.

7.3.3 Forgetting and Other Memory Problems

Another potential source of error in survey responses is failure to remember relevant information. Sometimes, respondents cannot remember the relevant events at all, and sometimes they remember the events only sketchily or inaccurately. It is useful to distinguish several forms of memory failure, since they can have different effects on the final answers:

1) Mismatches between the terms used in the question and the terms used to encode the events initially
2) Distortions in the representation of the events over time
3) Retrieval failure
4) Reconstruction errors

The first type of memory failure occurs when the terms the respondent uses to encode an event differ so markedly from the terms used in the question that the question does not call to mind the intended memories. For example, a respondent may not think of a glass of wine with dinner as an "alcoholic beverage." As a result, an item that asks about weekly consumption of alcoholic beverages may fail to trigger the relevant memories. Similarly, most of us probably do not think of trips to the hardware store as "shopping"; thus, it is very important that the NCVS item on shopping explicitly mentions hardware stores ("On average, during the last 6 months, that is, since _____, how often have YOU gone shopping? For example, at drug, clothing, grocery, hardware, and convenience stores"). When survey researchers develop questionnaires, they often conduct focus groups to learn how potential survey respondents think and talk about the survey topic. Both comprehension and retrieval improve when the terms used in

the questionnaire match those used by the respondents in encoding the relevant events.

A second source of inaccuracy in memory is the addition of details to our representation of an event over time. Most autobiographical memories probably consist of a blend of information we took in initially, while or shortly after we experienced the event, and information we added later on, as we were recounting the event to others who were not there themselves, reminiscing with others who also experienced it, or simply thinking about it later on. For example, when we think back to our high school graduation, our memory includes information we took in at the time, plus information we added later in looking at yearbooks, photographs, or videos of the event. These activities, which memory researchers term "rehearsal," play a pivotal role in maintaining vivid memories (e.g., Pillemer, 1984; Rubin and Kozin, 1984). Unfortunately, it is very difficult for us to identify the source of the information we remember; we cannot always distinguish what we experienced firsthand from what we merely heard about or inferred after the fact. Thus, any distortions or embellishments introduced as we recount our experiences or share reminiscences may be impossible to separate from the information we encoded initially. This sort of "postevent information" is not necessarily inaccurate, but it can be and once it is woven into our representation of an event it is very difficult to get rid of.

rehearsal

Still another source of trouble is retrieval failure—the failure to bring to mind information stored in long-term memory. We already noted one reason for retrieval failure: the question may not trigger recall for an event because it uses terms that differ too much from those used in encoding the event. Another reason for retrieval failure is the tendency to lose one memory among the other memories for similar experiences. Over time, it gets increasingly difficult to remember the details that distinguish one event, say, a specific doctor visit, from other events of the same kind; instead, the events blur together into a "generic memory" for a typical doctor visit, trip to the store, business trip, or whatever (cf. Barsalou, 1988; Linton, 1982). The accumulation of similar events over time means that we have more difficulty remembering specific events as more time elapses. The impact of the passage of time is probably the strongest and most robust finding to emerge from more than 100 years of research on forgetting (Rubin and Wetzel, 1996). Although researchers are still unclear about the exact shape of the function relating forgetting to the passage of time, it is clear that forgetting is more rapid at first and levels off thereafter. The amount forgotten in a given period also depends on the type of event in question; one study shows that people could still remember the names of nearly half of their classmates 50 years later (Bahrick, Bahrick, and Wittlinger, 1975). Figure 7.2 shows the percent of events correctly recalled for different kinds of phenomena. Although classmates are relatively easy to recall, the decay rate over time for recall of grades is very high. The best antidotes to retrieval failure in surveys seem to be providing more retrieval cues and getting the respondents to spend more time trying to remember. Table 7.1 (adapted from Tourangeau, Rips, and Rasinski, 2000) provides a more comprehensive list of the factors affecting recall and their implications for survey design. In short, the events most easily recalled are recent, distinctive, near another easily recalled event, and important in the life of the respondent. Questions that work best have rich, relevant cues and give the respondent time and encouragement to think carefully.

retrieval failure

generic memory

This tendency for recall and reporting to decline as a function of length of recall has yielded an important measurement error model. In the model μ_i is the

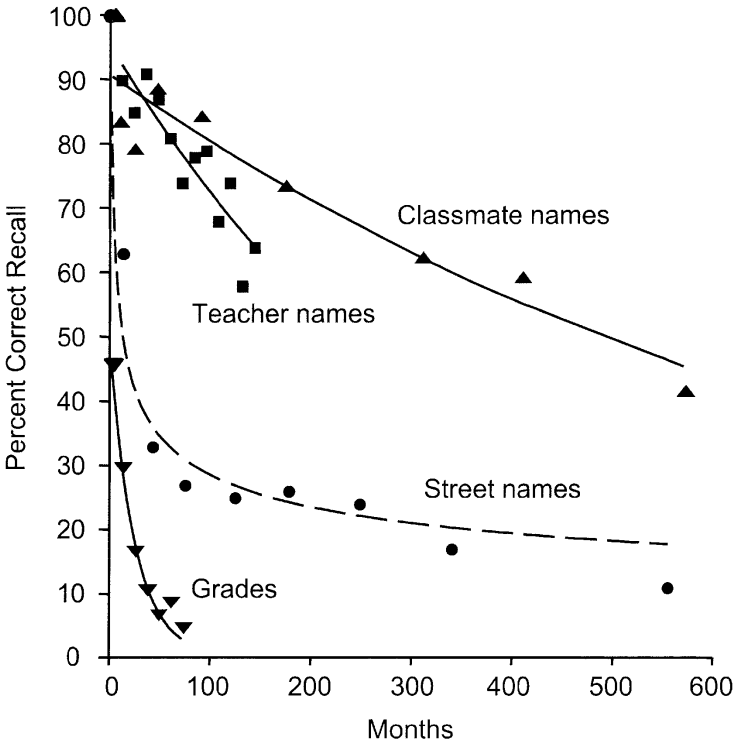

Figure 7.2 Recall accuracy for types of personal information. (Source: Tourangeau, Rips, and Rasinski, 2000.)

number of events experienced by the ith respondent relevant to the survey question. That is, if there were no recall problems, the ith person would report μ_i in response to the question. The model specifies that instead of μ_i, the respondent reports y_i:

$$y_i = \mu_i(ae^{-bt}) + \varepsilon_i$$

where a is the proportion of events that are reported (reflecting concerns about sensitivity or social desirability), b is the rate of decline in reporting as a function of time, and ε_i is a deviation from the model for the ith respondent. The e is just Euler's number, the base of the natural logarithms. Thus, the model specifies that the proportion of events correctly reported exponentially declines (rapid decline in the time segments immediately prior to the interview and diminishing declines far in the past). The literature implies that for events that are distinctive, near an easily recalled temporal boundary, and important in the life of the respondent, a is close to 1.0 and b is close to 0.0. For nonsensitive events that are easily forgotten, a may be close to 1.0, but b is large. As seen in Figure 7.2, the exponential decay model fits some empirical data better than others.

reconstruction

reference period

A final form of memory failure results from our efforts to reconstruct, or fill in, the missing pieces of incomplete memories. Such reconstructions are often based on what usually happens or what is happening right now. For instance,

Table 7.1. Summary of Factors Affecting Recall

Variable	Finding	Implication for Survey Design
Characteristics of Event		
Time of occurrence	Events that happened long ago are harder to recall	Shorten the reference period
Proximity to temporal boundaries	Events near significant temporal boundaries are easier to recall	Use personal landmarks, or life events calendars to promote recall
Distinctiveness	Distinctive events are easier to recall	Tailor the length of the reference period to the target events; use multiple cues to single out individual events
Importance, emotional impact	Important, emotionally involving events are easier to recall	Tailor the length of reference period to the properties of the target events
Question Characteristics		
Recall order	Backward search may promote fuller recall	Not clear whether backward recall is better in surveys
Number and type of cues	Multiple cues are typically better than a single cue; cues about the type of event (what) are better cues about participants or location (who or where), which are better than cues about when they occured	Provide multiple cues; use decomposition
Time on task	Taking more time improves recall	Use longer introductions to questions; slow the pace of the interview

Smith's studies of dietary recall suggested that respondents filled in gaps in their memories for what they actually ate with guesses based on what they usually eat. In a classic study, Bem and McConnell (1974) demonstrated a different strategy. Respondents in that study inferred their past views from what they now thought about the issue. This "retrospective" bias has been replicated several times (e.g., Smith, 1984). Our current state influences our recollection of the past with other types of memory as well, such as our recall of pain, past use of illicit substances, or income in the prior year (see Pearson, Ross, and Dawes, 1992 for further examples). We seem to reconstruct the past by examining the present and projecting it backwards, implicitly assuming that the characteristic or behavior in question is stable. On the other hand, when we remember that there has been a change, we may exaggerate the amount of change.

Many survey items ask respondents to report about events that occurred within a certain time frame or reference period. The first three of our sample items all specify a reference period extending from the moment of the interview to a specific date in the past (such as the date exactly six months before). Such questions assume that respondents are reasonably accurate at placing events in time. Unfortunately, dates are probably the aspect of events that are hardest for people to remember with any precision (e.g., Wagenaar, 1986). Although with some events—birthdays, weddings, and the like—people clearly do encode the date, for most events, the date is not something we are likely to note and remember. Because of the resulting difficulty in dating events, respondents may make "telescoping" errors in which they erroneously report events that actually occurred before the beginning of the reference period. The term "telescoping" suggests that past events seem closer to the present than they are; actually, recent

Neter and Waksberg (1964) on Response Errors

In 1964, Neter and Waksberg published a study comparing different designs for reporting past events.

Study design: Two design features were systematically varied: whether the interview was bounded (i.e., the respondents were reminded about their reports from the prior interview) and the length of the recall period (i.e., 1, 3, or 6 months). The context was a survey of the number of residential repair and renovation jobs and the expenditures associated with them, using household reporters.

Findings: With unbounded interviews, there were much higher reports of expenditures than with bounded interviews (a 55% increase). The increase in reports was larger for large jobs. The authors conclude that respondents were including reports of events that occurred before the reference period (this was labeled "forward telescoping"), and that rare events were subject to greater telescoping. Asking people to report events 6 months earlier versus 1 month earlier led to lower reports per month, with smaller jobs being disproportionately dropped from the longer reference periods. For the 6 month reference periods, the number of small jobs was 32% lower per month than for the 1 month reference period. The authors conclude this is a combined effect of failure to report and forward telescoping.

Limitations of the study: There were no independent data on the jobs or expenditures. Hence, the authors based their conclusions on the assumption that the bounded interviews offer the best estimates. Some respondents had been interviewed multiple times and may have exhibited different reporting behaviors.

Impact of the study: This study greatly sensitized designers to concerns about length of reference periods on the quality of reports. It encouraged the use of bounding interviews, for example, as in the National Crime Victimization Survey.

telescoping

studies suggest that "backward" telescoping is also common. As more time passes, we make larger errors (in both directions) in dating events (Rubin and Baddeley, 1989).

Despite this, telescoping errors tend, on the whole, to lead to overreporting. For example, in one classic study (see box on page 233), almost 40% of the home repair jobs reported by the respondents were reported in error due to telescoping (Neter and Waksberg, 1964). A procedure called "bounding" is sometimes used to reduce telescoping errors in longitudinal surveys. In a bounded interview, the interviewer reviews with the respondent a summary of the events the respondent reported in the previous interview. This is the procedure that was used in the NCVS until 2007 to attempt to eliminate duplicate reports of an incident reported in an earlier interview. The first of seven NCVS interviews asked about victimization incidents in the last six months, but data from this interview were not used in NCVS estimates. Instead, the second interview used first interview incident reports to "bound" the second interview reports. The interviewer asked whether an incident reported in the second interview might be a duplicate report of one in the first interview by checking the first interview reports. Similar procedures are used in later waves, always using the immediately prior interview as a "bound." This procedure sharply reduces the chance that respondents will report the same events in the current interview due to telescoping. Starting in 2007, because of cost pressures on the agency, the first interview data began to be used, with a statistical adjustment, in the annual NCVS estimates.

bounding

7.3.4 Estimation Processes for Behavioral Questions

Depending on what they recall, respondents may be forced to make an estimate in answering a behavioral frequency question or a render a judgment in answering an attitude question. Consider two of the examples we gave earlier in the chapter:

> Now turning to business conditions in the country as a whole – do you think that during the next 12 months we'll have good times financially, or bad times, or what? [SOC]

> Now think about the past 12 months, from [DATE] through today. We want to know how many days you've used any prescription tranquilizer that was not prescribed to you or that you took only for the experience or feeling it caused during the past 12 months. [NSDUH]

Some respondents may have a preexisting judgment about the economy that they are ready to report in response to a question like that in the SOC, but most respondents probably have to put together a judgment on the fly. Similarly, only a few respondents are likely to keep a running tally of the times they have abused prescription tranquilizers; the rest must come up with a total through some estimation process. (It is useful to note that NSDUH permits the respondent to answer the question with different metrics: average days per week, average days per month, or total days.) With both attitude and behavioral questions, the need

for respondents to put together judgments on the spot can lead to errors in the answers.

Let us first examine behavioral frequency questions like those in the NSDUH. Besides recalling an exact tally, respondents make use of three main estimation strategies to answer such questions:

recall-and-count

rate-based estimation

impression-based estimation

1) They may remember specific incidents and total them up, perhaps adjusting the answer upward to allow for forgotten incidents ("recall-and-count").
2) They may recall the rate at which incidents typically occur and extrapolate over the reference period ("rate-based estimation").
3) They may start with a vague impression and translate this into a number ("impression-based estimation").

For example, one respondent may recall three specific occasions on which he used prescription tranquilizers and report "3" as his answer. Another respondent may recall abusing prescription tranquilizers roughly once a month over the last year and report "12" as her answer. A third respondent may simply recall that he used the drugs a "few times" and report "5" as the answer. The different strategies for coming up with an answer are prone to different reporting problems.

The recall-and-count strategy is prone both to omissions due to forgetting and false reports due to telescoping. Depending on the balance between these two sources of error, respondents may systematically report fewer incidents than they should have or too many. Generally, the more events there are to report, the lower the accuracy of answers based on the recall-and-count strategy; with more events, it is both harder to remember them all and harder to total them up mentally. As a result, respondents tend to switch to other strategies as the reference period gets longer and as they have more incidents to report (Blair and Burton, 1987; Burton and Blair, 1991).

Several studies have asked respondents how they arrived at their answers to behavioral frequency questions like the ones in the NSDUH, and the reported popularity of the recall-and-count strategy falls sharply as the number of events to recall increases. Instead, respondents often turn to rate-based estimation when there are more than seven events or so to report. The literature suggests that rate-based estimation often leads to overestimates of behavioral frequencies. Apparently, people overestimate rates when the rate fluctuates or when there are exceptions to what usually happens.

But the most error-prone strategies for behavioral frequency questions are those based on impressions. When the question uses a closed format, the response options that it lists can affect impression-based estimates. When the answer categories emphasize the low end of the range, the answers tend to be correspondingly lower; when they emphasize the high end of the range, the answers tend to be high. The box on page 236 shows the results of a study that asked respondents how much television they watched in a typical day. Depending on which set of answer categories they got, either 16.2% or

Overreporting and Underreporting

When respondents report things that did not happen at all or report more events than actually occurred, this is called "overreporting." Certain types of things are characteristically overreported in surveys. For example, in any given election, more people say that they voted than actually did. The opposite error is called "underreporting" and involves reporting fewer events than actually took place.

Schwarz, Hippler, Deutsch, and Strack (1985) on Response Scale Effects

In 1985, Schwarz, Hippler, Deutsch, and Strack published the results of several studies measuring the effects of response scales on reporting.

Study design: Two different randomized experiments were embedded in larger surveys. A between-subject design administered one form of a question to one-half of the sample, and another, to the other. One experiment used a quota sample of 132 adults; the other, 79 employees of an office recruited into a survey. One-half of the sample received a question about hours spent watching television, with a six-category scale with middle categories 1–1.5 hours and 1.5–2 hours; for the other half-sample, the middle categories were 3–3.5 hours and 3.5–4 hours.

Low Options		High Options	
Responses	%	Responses	%
< ½ hr	7.4%	< 2½ hr	62.5%
½ to 1 hr	17.7%	2½ to 3 hr	23.4%
1 to 1½ hr	26.5%	3 to 3½ hr	7.8%
1½ to 2 hr	14.7%	3½ to 4 hr	4.7%
2 to 2½ hr	17.7%	4 to 4½ hr	1.6%
>2½ hr	16.2%	> 4½ hr	0.0%
Total	100.0%		100.0%

Findings: Respondents receiving the low average scale tended to report watching less television than those receiving the high average scale. For example, 16.2% reported watching more than 2.5 hours per day in the low average scale but 37.5% in the high average scale.

Limitations of the study: One of the studies used a quota sample, the other, a group of office workers, limiting the ability to generalize to other survey conditions.

Impact of the study: The studies demonstrated that response scales affect reporting of behaviors. Researchers now attempt to choose center categories that are closest to the expected population averages.

37.5% of the respondents said they watched more that two-and-a-half hours per day. Impression-based estimates are also prone to wild values when the question is posed in open-ended format.

7.3.5 Judgment Processes for Attitude Questions

Responding to an attitude question might seem to involve very different cognitive processes from those needed to answer factual items about behavior, but at least some authors (e.g., Tourangeau, Rips, and Rasinski, 2000) argue that there are more similarities than differences between the response processes for the two types of items. The same four types of information from which frequency estimates are derived—exact tallies, impressions, generic information, and specific memories—have their counterparts in attitude questions. For instance, some respondents (economists and people who follow the stock market) may have clearly defined views on which to base an answer to the item about the economy from the Survey of Consumers ("Do you think that during the next 12 months we'll have good times financially, or bad times, or what?"). Others may have a vague impression ("Gee, I read something in the *Wall Street Journal* the other day and it sounded pretty ominous"). Just as we may have only a hazy sense of how often we have done something, we may have an equally vague impression of a person or issue we were asked to evaluate. Or, lacking any ready-made evaluation (even a very hazy one), we may attempt to construct one either from the top down, deriving a position from more general values or predispositions, or from the bottom up, using specific beliefs about the issue to construct an opinion about it. The latter two strategies resemble the use of generic information and the recall-and-count strategy to answer frequency questions.

When respondents do not have an existing evaluation they can draw on, their answers to an attitude question may be strongly affected by the exact wording of the question or by the surrounding context in which the question is placed. Consider these two items, both administered in the early 1950s to gauge public support for the Korean War:

> Do you think the United States made a mistake in deciding to defend Korea or not? [Gallup]
>
> Do you think the United States was right or wrong in sending American troops to stop the Communist invasion of South Korea? [NORC]

The NORC item consistently showed higher levels of support for the Korean War than the Gallup item did. In a series of experiments, Schuman and Presser (1981) later showed that adding the phrase "to stop a Communist takeover" increased support for U.S. military interventions by about 15 percentage points. Several studies have shown similar wording effects on other topics; there is far more support for increased spending on halting the rising crime rate than on law enforcement, for aid to the poor than for welfare, for dealing with drug addiction than with drug rehabilitation, and so on (Rasinski, 1989; Smith, 1987). The wording of an item can help (and influence) respondents who need to infer their views on the specific issue from more general values; the NORC wording apparently reminded some respondents that the general issue was the spread of Communism and that helped them formulate their judgment about the U.S. role in Korea.

Question context can also shape how respondents evaluate an issue. Most attitude judgments are made on a relative basis. When we evaluate a political figure, say, that evaluation almost inevitably involves comparisons to rival candidates, to other salient political figures, or to our image of the typical politician. The standard of comparison for the judgment is likely to have an impact on which characteristics of the political figure come to mind and, more importantly, on how those characteristics are evaluated. A Democrat may evaluate Bill Clinton's terms in office quite favorably when the standard of comparison is the Reagan administration but less favorably when the standard is that of Franklin Delano Roosevelt.

7.3.6 Formatting the Answer

Once they have generated an estimate or an initial judgment, respondents confront a new problem—translating that judgment into an acceptable format. Survey items can take a variety of formats, and we focus on the three most common:

open and closed questions

1) Open-ended questions that call for numerical answers
2) Closed questions with ordered response scales
3) Closed questions with categorical response options

The examples below, taken from the BRFSS, illustrate each of these formats.

1) Now, thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 was your physical health not good?

2) Would you say that in general your health is:
   1 Excellent
   2 Very good
   3 Good
   4 Fair
   5 Poor

3) Are you:
   1 Married
   2 Divorced
   3 Widowed
   4 Separated
   5 Never married
   6 A member of an unmarried couple

For formats (2) and (3), the interviewer is instructed to "please read" the answer categories (but not the numbers attached to them). Almost all of the items in the BRFSS follow one of these three formats or ask yes–no questions. Yes–no questions are probably closest to closed categorical questions. The BRFSS is not unusual in relying almost exclusively on these response formats; most other surveys do so as well.

Each of the three formats presents their own special challenges to respondents. With numerical open-ended items like those in (1), respondents may have a lot of trouble translating a vague underlying judgment ("I had a pretty bad month") into an exact number ("I felt sick on three days"). There are good reasons why open-ended items are popular with survey researchers. In principle, open-ended items yield more exact information than closed items. Even with finely graded response options, there is inevitably some loss of information when the answer is categorical. Moreover, the answer categories must often be truncated at the high or low end of the range. For example, at one point the BRFSS included an open item asking respondents how many sex partners they have had in the last 12 months. A closed item would have to offer some top category (e.g., "10 or more") that would yield inexact information about the most sexually active part of the population. So relative to open items, closed items lose information because of the grouping and truncation of possible answers.

In practice, however, respondents often seem to act as if open questions were not really seeking exact quantities. Tourangeau and his colleagues, for instance, found that most of the respondents in a survey of sexual behavior who reported 10 or more sexual partners gave answers that were exact multiples of five (Tourangeau, Rasinski, Jobe, Smith, and Pratt, 1997). Survey respondents also report many other quantities as round values, such as how much stress they feel in caring for relatives with disabilities (Schaeffer and Bradburn, 1989) or how long ago they completed an earlier interview (Huttenlocher, Hedges, and Bradburn, 1990). With items asking for percentages, the responses tend to cluster at 0, 50, and 100. Although several things contribute to the use of round numbers (such as fuzziness about the underlying quantity or embarrassment about the

value to be reported), the main problem seems to be the sheer difficulty many respondents experience in assigning numbers to their estimates and judgments. Respondents may try to simplify this task by selecting a value from a limited number of ranges; they then report the range they chose as a round value.

Scale ratings, such as item (2) above, also have their characteristic problems. With some types of ratings (e.g., personnel ratings), respondents seem to shy away from the negative end of the scale, producing "positivity bias." With other types of rating scales, respondents tend to avoid the most extreme answer categories. When the scale points have numerical labels, the labels can affect the answers. A study by Schwarz and his collaborators (Schwarz, Knäuper, Hippler, Noelle-Neumann, and Clark, 1991) illustrates this problem. They asked respondents to rate their success in life. One group of respondents used a scale that ranged from –5 to +5; the other group used a scale that ranged from 0 to 10. In both cases, the end points of the scale had the same verbal labels. With both sets of numerical labels, the ratings tend to fall on the positive half of the scale (exhibiting an overall positivity bias), but the heaping on the more positive end of the scale was more marked when the scale labels ran from –5 to +5. Negative numbers, according to Schwarz and his collaborators, convey a different meaning from numbers ranging from 0 upward. A value of 0 implies a lack of success in life, whereas a value of –5 implies abject failure.

positivity bias

At least two other features of rating scales can affect the answers: the labeling and number of options presented (e.g., five points vs. nine points). Krosnick and Berent (1993) conducted a series of studies comparing two types of response scales: ones that labeled only the end categories and ones that provided verbal labels for every category. In addition, they compared typical rating scale items, which present all the response options in a single step, to branching items, which present a preliminary choice ("Are you a Republican, Democrat, or Independent?") followed by questions offering more refined categories ("Are you strongly or weakly Democratic?"). Both labeling and the two-step branching format increased the reliability of the answers. Krosnick and Berent argue that the labels help clarify the meaning of the scale points and that the branching structure makes the reporting task easier by breaking it down into two simpler judgments. Aside from labeling and branching, the sheer number of the response categories can affect the difficulty of the item. With too few categories, the rating scales may fail to discriminate between respondents with different underlying judgments; with too many, respondents may fail to distinguish reliably between adjacent categories. Krosnick and Fabrigar (1997) argue that seven scale points seems to represent the best compromise.

The final common response format offers respondents unordered response categories [like the marital status item given in (3) above]. One problem with this format is that respondents may not wait to hear or read all of the options; instead, they may select the first reasonable answer they consider. For example, with the BRFSS item on marital status a respondent may select the "Never married" option without realizing that the final option ("Member of an unmarried couple") offers a better description of his or her situation. Several studies have compared what happens when the order of the response options is reversed. These studies have found two types of effects: primacy and recency effects. With a "primacy effect," presenting an option first (or at least near the beginning of the list) increases the chances that respondents will choose that option. With a "recency effect," the opposite happens—putting an option at or near the end of the list increases its

primacy effect

recency effect

popularity. Most researchers believe that respondents are likely to consider the response options one at a time and to select the first one that seems to provide an adequate answer. Using the terminology of Krosnick (1999), respondents satisfice rather than optimize; they pick an answer that is good enough, not necessarily the one that is best. The tendency to take such shortcuts would explain why primacy effects are common. The reason that recency effects also occur is that respondents may not consider the answer options in the same order as the questionnaire presents them. When an interviewer reads the questions to the respondent, the last option the interviewer reads may be the first one that respondents think about. (By contrast, when respondents read the questions themselves, they are more likely to read and consider the response categories in order.) Because of this difference, respondents in telephone surveys seem prone to recency effects, whereas respondents in mail surveys are more prone to primacy effects. Response order effects are not just a survey phenomenon; Krosnick has shown that the order in which candidates for political office are listed on the ballot affects the share of the vote each one gets.

7.3.7 Motivated Misreporting

So far, we have focused on respondents' efforts to deal with the cognitive difficulties posed by survey questions, but a glance at many survey questionnaires reveals many difficulties of another sort. Consider these items from the NSDUH:

a) Think specifically about the last 30 days, from _____ up to and including today. During the past 30 days, on how many days did you use cocaine?
c) How long has it been since you last smoked part or all of a cigarette?

It is easy to imagine someone having little difficulty in interpreting (a) or in retrieving and formatting the information it asks for, but still giving the wrong answer. Survey researchers call questions like these "sensitive" or "threatening," and they have become more common on national surveys as researchers attempt to monitor the use of illicit drugs or the behaviors that contribute to the spread of HIV/AIDS. "Sensitive questions" are questions that are likely to be seen as intrusive or embarrassing by some respondents. For example, questions about personal income or about sexual behavior fall in this category. Respondents are more likely to refuse to answer such questions or to give deliberately wrong answers to them. Sensitive questions create a dilemma for respondents; they have agreed to help the researchers out by providing information, but they may be unwilling to provide the information that specific questions are requesting. Respondents often seem to resolve such conflicts by skipping the questions or providing false answers to them. For instance, one study (Moore, Stinson, and Welniak, 1997) has looked at missing data in questions about income. Moore and his colleagues report that more than a quarter of the wage and salary data in the Current Population Survey (CPS) is missing or incomplete. This is roughly ten times the rate of missing data for routine demographic items.

sensitive question

Sometimes, refusing to answer a question may be more awkward than simply underreporting some embarrassing behavior. For example, refusing to answer

the NSDUH item about cocaine use [example (a)] is tantamount to admitting one has used cocaine. It may be easier just to deny using cocaine. Among the potentially embarrassing behaviors that seem to be underreported in surveys are the use of illicit drugs, the consumption of alcohol, smoking (especially among teens and pregnant women), and abortion. Respondents may also be reluctant to admit that they have not done something when they feel they should have. As a result, they may overreport certain socially desirable behaviors such as voting or going to church.

Some researchers have attempted to use "forgiving" wording to improve the reporting of sensitive information. Consider, for example, this question about voting:

> In talking to people about elections, we often find that a lot of people were not able to vote because they were not registered, they were sick, or they just didn't have the time. How about you—did you vote in the elections this November? (American National Election Study)

The wording of the voting question encourages respondents to report that they did not vote. Despite this, the findings suggest that such wording does not eliminate overreports of voting. The most important single tactic for improving reports on sensitive topics seems to be removing the interviewer from the question-and-answer process (see Section 5.3.5). This can be accomplished in several ways. First, the sensitive questions can be administered on a paper self-administered questionnaire or presented directly to the respondent by a computer. Both of these forms of self-administration seem to increase reporting of potentially embarrassing information compared to interviewer administration of the questions. Another technique is called the "randomized response technique" (Warner, 1965). With this method, respondents spin a dial or use some other chance device to determine whether they answer the sensitive question or a second, innocuous question ("Were you born in September?"). People seem to be more willing to answer truthfully when the interviewer does not know which question they are answering. Estimates of the prevalence of the sensitive characteristics are based on the known probabilities of the randomizing device assigning the sensitive question and the nonsensitive question. Evaluations of the technique in practice suggest that it reduces some but not all of the bias common in answering sensitive questions.

randomized response technique

7.3.8 Navigational Errors

When the questions are self-administered (for example, in a mail questionnaire), the respondents have to understand both the questions themselves and any instructions the questionnaire includes, about which questions they are supposed to answer, what form their answers should take (for example, "Mark one"), and any other instructions. In fact, in a self-administered questionnaire, an important part of the respondents' job is to figure out where to go next after they have answered a question. To help the respondents find the right path through the questionnaire, questionnaires often include various skip instructions ("If No, go to Question 8"); these verbal instructions are often reinforced by visual and graphical cues, such as boldfacing and arrows. Figure 7.3 is an example adapted from

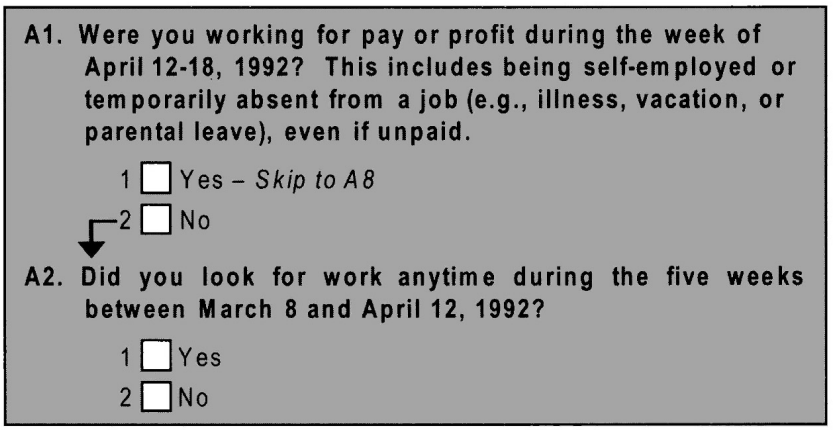

Figure 7.3 Example questions from Jenkins and Dillman (1997).

one discussed by Jenkins and Dillman (1997). This example illustrates several principles that help respondents find their way through questionnaires. Throughout the questionnaire, fonts, boldface, and graphical symbols are used in a consistent way. For example, the questions themselves are set off from other text by the use of bold type; the question numbers (A1 and A2) are made to stand out even more prominently by their position at the extreme left margin. The routing instructions (*Skip to A8*) for the first item (and for subsequent items) are in italics. Spaces that the respondents are supposed to fill in are in white, a color that contrasts with the shaded background. In addition, where possible, arrows rather than verbal instructions are used to convey the intended path (e.g., those who answer "No" to Question A1 are directed to A2 by an arrow).

navigational error

Still, it is easy for respondents to make "navigational errors," to skip items they were supposed to answer or to answer questions they were supposed to skip. Respondents may not notice instructions or, if they do, they may not understand them. As a result, self-administered questionnaires, especially poorly designed ones, often have a higher rate of missing data than interviewer-administered questions do.

With the increase in self-administered questionnaires in ACASI, Web, and e-mail surveys, there is much methodological research needed to discover how respondents react to alternative formats. What properties of formats reduce the burden on respondents? Do low-literacy respondents profit more from different formats than high-literacy respondents? Can formatting act to increase the motivation of respondents to perform their tasks?

7.4 Guidelines for Writing Good Questions

There is value in being aware of all the potential pitfalls in questionnaire design; it makes it easier to recognize the problems with a question but it is also useful to have some positive rules for avoiding these problems in the first place. Several textbooks provide guidelines for writing good survey questions, and this section summarizes one of the most comprehensive lists, the one developed by Sudman and Bradburn (1982). (Another source of very good advice about writing survey

questions is Converse and Presser, 1986.) Where we believe Sudman and Bradburn's recommendations have not stood the test of time, we drop or amend their original advice. Their recommendations were based empirical findings and, for the most part, they have held up pretty well. Many of them have already been foreshadowed in our discussion of the things that can go wrong in the response process.

Sudman and Bradburn do not offer a single set of guidelines for all survey questions but instead give separate recommendations for several different types of questions:

1) Nonsensitive questions about behavior
2) Sensitive questions about behavior
3) Attitude questions

These distinctions are useful, since the different types of question raise somewhat different issues. For example, sensitive questions are especially prone to deliberate misreporting and may require special steps to elicit accurate answers. Attitude questions are likely to involve response scales, and response scales (as we noted earlier) raise their own special problems. We will deal with each type of question in turn.

7.4.1 Nonsensitive Questions About Behavior

With many questions about behavior, the key problems are that respondents may forget some or all of the relevant information or that their answers may reflect inaccurate estimates. Accordingly, many of the Sudman and Bradburn's guidelines for nonsensitive questions about behavior are attempts to reduce memory problems. Most of their guidelines for nonsensitive questions make equally good sense for sensitive ones as well. Here they are:

1) With closed questions, include all reasonable possibilities as explicit response options.
2) Make the questions as specific as possible.
3) Use words that virtually all respondents will understand.
4) Lengthen the questions by adding memory cues to improve recall.
5) When forgetting is likely, use aided recall.
6) When the events of interest are frequent but not very involving, have respondents keep a diary.
7) When long recall periods must be used, use a life event calendar to improve reporting.
8) To reduce telescoping errors, ask respondents to use household records or use bounded recall (or do both).
9) If cost is a factor, consider whether proxies might be able to provide accurate information.

The first three recommendations all concern the wording of the question. It is essential to include all the possibilities in the response categories because

respondents are reluctant to volunteer answers that are not explicitly offered to them. In addition, possibilities that are lumped together in a residual category ("All Others") tend to be underreported. For example, the two items below are likely to produce very different distributions of answers:

1) What is your race?
   White
   Black
   Asian or Pacific Islander
   American Indian or Alaska Native
   Some Other Race

2) What is your race?
   White
   Black
   Asian Indian
   Chinese
   Japanese
   Korean
   Vietnamese
   Filipino
   Other Asian
   Native Hawaiian
   Guamanian
   Samoan
   Some Other Pacific Islander
   American Indian or Alaska Native
   Some Other Race

Unpacking the "Asian and Pacific Islander" option into its components clarifies the meaning of that answer category and also makes it easier for respondents to recognize whether it is the appropriate option for them. As a result, the second item is likely to yield a higher percentage of reported Asian and Pacific Islanders.

Making the question as specific as possible reduces the chances for differences in interpretation across respondents. It is important that the question be clear about who it covers (does "you" mean just the respondent or everyone in the respondent's household?), what time period, which behaviors, and so on. A common error is to be vague about the reference period that the question covers: "In a typical week, how often do you usually have dessert?" Our eating habits can vary markedly over the life course, the year, even over the last few weeks. A better question would specify the reference period: "Over the last month, that is, since [DATE], how often did you have dessert in a typical week?" The interviewer would fill in the exact start date at the time he or she asked the question.

The third recommendation is to use words that everyone understands, advice that is unfortunately far easier to give than to follow. Some more specific guidelines are to avoid technical terms ("Have you ever had a myocardial infarction?") in favor of everyday terms ("Have you ever had a heart attack?"), vague quanti-

fiers ("Often," "Hardly ever") in favor of explicit frequency categories ("Every day," "Once a month"), and vague modifiers ("Usually") in favor of more concrete ones ("Most of the time"). If need be, vague or technical terms can be used but they should be defined, ideally just before the question ("A myocardial infarction is a heart attack; technically, it means that some of the tissue in the heart muscle dies. Have you ever had a myocardial infarction?").

The next five guidelines are all about reducing the impact of forgetting on the accuracy of survey reports. One basic strategy is to provide more retrieval cues to the respondent, either by incorporating them into the question itself or by asking separate questions about subcategories of the overall category of interest. Adding cues may lengthen the question, but, as the fourth guideline notes, it may improve recall as well. Asking separate questions about subcategories is referred to as "aided recall," and this is what the fifth guideline recommends. An example is the NCVS item on shopping, which lists several examples of different places one might shop ("... drug, clothing, grocery, hardware, and convenience stores"); it would also be possible to ask separate questions about each type of retail outlet. Both approaches provide retrieval cues that help jog respondents' memories. It is important that the retrieval cues are actually helpful. Breaking a category down into nonsensical subcategories ("How many times did you go shopping for red things?" "How many times did you go shopping on a rainy Tuesday morning?") can make things worse (see Belli, Schwarz, Singer, and Talarico, 2000). Retrieval cues that do not match how the respondents encode the events can be worse than no cues at all. Another strategy for improving recall is to tailor the length of the reference period to the likely memorability of the events. Things that rarely happen, those that have a major emotional impact, and those that last a long time tend to be easier to remember than frequent, inconsequential, or fleeting events. For example, respondents are more likely to remember a hospital stay for open heart surgery than a 15-minute visit to the doctor for a flu shot. As a result, it makes sense to use a longer reference period to collect information about memorable events, such as hospital stays (where one year might be a reasonable reference period), than about nonmemorable events, such as doctor visits (where two weeks or a month might be reasonable).

aided recall

The tailoring approach has its limits, though. When the survey concerns very routine and uninvolving events (say, small consumer purchases or food intake), the reference period might have to be too short to be practical, say, yesterday. In such cases, it is often better to have respondents keep a diary rather than rely on their ability to recall the events. At the other extreme, a long reference period may be a necessary feature of the survey design. Most panel surveys, for example, cannot afford to visit the respondents very often; generally, they interview panel members every few months or even once a year and so must use long references periods if they intend to cover the whole period between interviews. (For example, the NCVS visits sample dwellings every six months.) When a long reference period has to be used, a life events calendar can sometimes improve recall. A life event calendar collects milestone events about several domains of a person's life, such as marital history, births of children, jobs, and residences. These are recorded on a calendar and help jog respondents' memories about more mundane matters, like how much they were earning at the time, illnesses they experienced, or crime victimizations. The autobiographical milestones recorded on the calendar provide rich chronological and thematic cues for retrieval (Belli, 1998); they can also serve as temporal landmarks, improving our ability to date other events.

These event calendars require the interviewer to engage in less structured interaction with the respondent. There are unanswered research questions about whether such tools increase variability in survey results from interviewer effects (see Section 9.3). This is a ripe area for methodological research.

Another kind of memory error involves misdating events (or "telescoping errors"), and the eighth guideline recommends two methods for reducing these errors. The first is to have respondents consult household records (calendars, checkbooks, bills, insurance forms, or other financial records) to help them recall and date purchases, doctor visits, or other relevant events that may leave some paper trail. The second tactic is called "bounding"; it involves reminding respondents what they already reported in a previous round of a panel survey.

proxy

The final recommendation is to use proxy reporters to provide information when data collection costs are an issue. A "proxy" is anyone other than the person about whom the information is being collected. Most surveys ask parents to provide information about young children rather than interviewing the children themselves. Other surveys use a single adult member of the household to report on everyone else who lives there. Allowing proxies to report can reduce costs, since the interviewer can collect the data right away from a proxy rather schedule a return trip to speak to every person in the household. At the same time, proxies differ systematically from self-reporters. They are, for example, more likely than self-reporters to rely on generic information (e.g., information about what usually happens) in answering the questions than on episodic information (e.g., detailed memories for specific incidents). In addition, self-reporters and proxies may differ in what they know. It hardly makes sense to ask parents about whether their teenaged children smoke; the children are likely to conceal this information, especially from their parents. On the whole, though, proxies often seem to provide reliable factual information (e.g., O'Muircheartaigh, 1991).

7.4.2 Sensitive Questions About Behavior

As we noted in Section 7.3.7, some surveys include questions about illegal or potentially embarrassing behaviors, such as cocaine use, drinking, or smoking. Both the NSDUH and the BRFSS are full of such questions. Here are Sudman and Bradburn's guidelines (updated as necessary) for sensitive questions:

1) Use open rather than closed questions for eliciting the frequency of sensitive behaviors.
2) Use long rather than short questions.
3) Use familiar words in describing sensitive behaviors.
4) Deliberately load the question to reduce misreporting.
5) Ask about long periods (such as one's entire lifetime) or periods from the distant past first in asking about sensitive behaviors.
6) Embed the sensitive question among other sensitive items to make it stand out less.
7) Use self-administration or some similar method to improve reporting.
8) Consider collecting the data in a diary.
9) At the end of the questionnaire, include some items to assess how sensitive the key behavioral questions were.
10) Collect validation data.

Open questions about sensitive behaviors have two advantages over closed questions. First, closed questions inevitably lose information (for example, about the very frequent end of the continuum). In addition, closed categories may be taken by the respondents as providing information about the distribution of the behavior in question in the general population and, thus, affect their answers (see the box describing the study by Schwarz and colleagues on page 236).

Sudman and Bradburn recommend longer questions, largely because they promote fuller recall (by giving respondents more time to remember) (Sudman and Bradburn, 1982); they are particularly useful with behaviors that tend to be underreported (such as drinking). For example, to ask about the consumption of wine, they recommend the following wording:

> Wines have become increasingly popular in this country in the last few years; by wines, we mean liqueurs, cordials, sherries, and similar drinks, as well as table wines, sparkling wines, and champagne. Did you ever drink, even once, wine or champagne?

The enumeration of the various types of wine may help clarify the boundaries of the category, but mostly serves to initiate and provide extra time for retrieval. The next recommendation, to use familiar terms for sensitive behaviors (e.g., "having sex" rather than "coitus"), may make respondents more comfortable with the questions but also tends to improve recall, since the terms in the question are more likely to match the terms used to encode the relevant experiences. Interviewers can determine which terms the respondents would prefer to use at the outset of the interview.

loading

"Loading" a question means wording it in a way that invites a particular response, in this case, the socially undesirable answer. Sudman and Bradburn distinguish several strategies for doing this (Sudman and Bradburn, 1982): the "everybody-does-it" approach ("Even the calmest parents get mad at their children sometimes. Did your children do anything in the past week to make you angry?"); the "assume-the-behavior" approach ("How many times during the past week did your children do something that made you angry?"); the "authorities-recommend-it" approach ("Many psychologists believe it is important for parents to express their pent-up frustrations. Did your children do anything in the past week to make you angry?"); and the "reasons-for-doing-it" approach ("Parents become angry because they're tired or distracted or when their children are unusually naughty. Did your children do anything in the past week to make you angry?"). The question about voting on page 241 also illustrates this last approach.

The next two recommendations help reduce the apparent sensitivity of the item. In general, it is less embarrassing to admit that one has ever done something or did it a long time ago than to admit one has done it recently. For example, during the 2000 presidential campaign, candidate Bush admitted he had had a drinking problem more than ten years before, and this admission provoked little reaction. It would have been quite a different story for him to admit he had been drinking heavily the day before on the campaign trail. Most survey researchers think that sensitive items should not come at the start of the interview, but only after some less sensitive questions. In addition, embedding one sensitive question (for example, an item on shoplifting) among other more sensitive items (an item on

armed robbery) may help make the sensitive item of interest seem less threatening by comparison. Like many judgments, the perception of sensitivity is affected by context.

As we already noted, one of the most effective methods for improving reports about sensitive behaviors is by having the respondent complete a self-administered or a computer-administered questionnaire. Another approach is the randomized response technique in which the interviewer does not know the question the respondent is answering. This method is illustrated below. Respondents pick a red or a blue bead from a box and their selection determines which question they answer:

(Red) Have you been arrested for drunk driving in the last 12 months?
(Blue) Is your birthday in the month of June?

The interviewer records either a "yes" or "no" answer, not knowing which question the respondent is answering. Since the researcher knows the probability of a red bead or a blue bead being selected (and the probability the respondent was born in June), an estimate of the proportion of "yes" answers to the drunk driving question can be obtained.

Finally, a third approach is having the respondents keep a diary, which combines the benefits of self-administration with the reduced burden on memory. Diary surveys that require detailed record keeping, however, tend to have lower response rates.

The final two recommendations allow us to assess the level of sensitivity of the questions (by having respondents rate their discomfort in answering them) and to assess the overall accuracy of responses by comparing them to an external benchmark. For example, respondents' survey reports about recent drug use might be compared to the results of a urinalysis.

7.4.3 Attitude Questions

Many surveys ask about respondents' attitudes. Among our six example surveys, only the SOC includes a large number of attitude items. Still, these are a very common class of survey questions, and Sudman and Bradburn present some guidelines specifically for them. Here is our amended version of their list:

1) Specify the attitude object clearly.
2) Avoid double-barreled questions.
3) Measure the strength of the attitude, if necessary using separate items for this purpose.
4) Use bipolar items except when they might miss key information.
5) The alternatives mentioned in the question have a big impact on the answers; carefully consider which alternatives to include.
6) In measuring change over time, ask the same questions each time.
7) When asking general and specific questions about a topic, ask the general question first.
8) When asking questions about multiple items, start with the least popular one.

9) Use closed questions for measuring attitudes.
10) Use five- to seven-point response scales and label every scale point.
11) Start with the end of the scale that is the least popular.
12) Use analogue devices (such as thermometers) to collect more detailed scale information.
13) Use ranking only if the respondents can see all the alternatives; otherwise, use paired comparisons.
14) Get ratings for every item of interest; do not use check-all-that-apply items.

The first six of these guidelines all deal with the wording of the questions. The first one says to clearly specify the attitude object of interest. Consider the item below:

> Do you think the government is spending too little, about the right amount, or too much on antiterrorism measures?

It would improve comprehension and make the interpretation of the question more consistent across respondents to spell out what antiterrorism measures the question has in mind (and what level of government). Double-barreled items inadvertently ask about two attitude objects at once. For example, question (a) below ties attitudes about abortion to attitudes about the Supreme Court and (b) ties abortion attitudes to attitudes about women's rights:

double-barreled items

a) The U.S. Supreme Court has ruled that a woman should be able to end a pregnancy at any time during the first three months. Do you favor or oppose this ruling?
b) Do you favor legalized abortion because it gives women the right to choose?

The answers to double-barreled items are difficult to interpret; do they reflect attitudes to the one or the other issue or both?

The two characteristics of an attitude that are generally of interest are its direction (agree or disagree, pro or con, favorable or unfavorable) and its intensity or strength. The third recommendation is to assess intensity, using a response scale designed to capture this dimension ("Strongly disagree," "Disagree somewhat," etc.), a separate item, or multiple items that can be combined into a scale that yields intensity scores. The fourth recommendation is to use bipolar items except where they might miss some subtle distinction. For example, ask about conflicting policies in a single item rather than asking about each policy alone:

> Should the government see to it that everyone has adequate medical care or should everyone see to his own medical care?

bipolar approach

The "bipolar approach" forces respondents to choose between plausible alternatives, thereby discouraging acquiescence. There are times, however, when this approach misses subtleties. For instance, positive and negative emotions are not

always strongly (negatively) related, and so it may make sense to include separate questions asking whether something makes respondents happy or sad. The fifth recommendation about the wording about attitude items concerns the consequences of including middle (e.g., "neither agree nor disagree") and no-opinion options. In general, these options should be included unless there is some compelling reason not to. (For example, in an election poll, it is important to get those leaning one way or the other to indicate their preferences; under those circumstances, middle and no-opinion options may be omitted.) The next guideline advises that the only way to measure changes in attitudes is to compare apples with apples, that is, to administer the same question at both time points.

The next two recommendations are designed to reduce the impact of question order. If a questionnaire includes both a general question and more specific questions about the same domain, it is probably best to ask the general item first; otherwise the answers to that item are likely to be affected by the number and content of the preceding specific items. (Recall our earlier discussion of how respondents reinterpret a question on overall happiness when it follows one asking about marital happiness.) If the questionnaire includes several parallel questions that vary in popularity (e.g., the GSS includes several similar items asking about support for abortion under different circumstances), the unpopular ones are likely to seem even less appealing when they follow the popular ones. Putting the unpopular items first may yield more revealing answers.

The final six recommendations concern the format of the response scales that are nearly ubiquitous with attitude items. The first of these recommendations (the ninth overall) advises us to use closed attitude items rather than open-ended ones. The latter are simply too difficult to code. The next recommendation gets more specific, suggesting that five to seven verbally labeled scale points be used. Fewer scale points lose information; more tend to produce cognitive overload. The verbal labels help ensure that all respondents interpret the scale in the same way. If the response options vary in popularity, more respondents will consider the less popular ones when they come first than if they come later. If interviewers administer the questions aloud, then respondents probably consider the last option they hear first; in such cases, the least popular option should come at the end.

analogue method

ranking

check-all-that-apply

Other formats are also popular for attitude questions. The last three recommendations concern analogue methods (such as feeling thermometers), rankings, and check-all-that-apply items. When more than seven scale points are needed, an analogue scale may help reduce the cognitive burden. For example, a feeling thermometer asks respondent to assess their warmth toward public figures using a scale that goes from 0 (indicating very cold feelings) to 100 (very warm). Results suggest that respondents are likely to use 13 or so points on the scale. Respondents have difficulty with unanchored numerical judgments (that is one reason they tend to use round numbers) so it helps when the scale has both an upper and a lower limit, as the feelings thermometer does. Respondents can also be asked to rank various objects (e.g., desirable qualities for a child to have). As the thirteenth guideline points out, the ranking task may exceed the cognitive capacity of the respondents unless all of the items to be ranked are displayed on a card that the respondents can look at while they rank the items. When that is impossible (e.g., in a telephone interview), the researchers may have to fall back on comparisons between pairs of objects. The final recommen-

dation discourages researchers from using check-all-that-apply items, since respondents are likely to check only some of the items that actually apply to them (Rasinski, Mingay, and Bradburn, 1994). Asking respondents to say yes or no (agree or disagree, favor or oppose, etc.) to each item on the list reduces this form of satisficing.

7.4.4 Self-Administered Questions

Sudman and Bradburn (and most other questionnaire design texts) focus on questionnaires for face-to-face and telephone interviews. In these settings, a trained interviewer typically mediates between the respondents and the questionnaire. By contrast, with mail questionnaires, it is up to the respondents to figure out which questions to answer, how to record their responses, and how to comply with any other instructions for completing the questionnaire. Jenkins and Dillman (1997; see also Redline and Dillman, 2002) offer several recommendations for improving the chances that respondents will correctly fill out mail and other self-administered questionnaires. Here are their recommendations:

1) Use visual elements in a consistent way to define the desired path through the questionnaire.
2) When the questionnaire must change its conventions part way through, prominent visual guides should alert respondents to the switch.
3) Place directions where they are to be used and where they can be seen.
4) Present information that needs to be used together in the same location.
5) Ask one question at a time.

The "visual elements" mentioned in the first guideline include brightness, color, shape, and position on the page. As we noted in Section 7.3.8, the questionnaire may set question numbers off in the left margin, put the question numbers and question text in boldface, put any instructions in a different typeface from the questions themselves, and use graphical symbols (such as arrows) to help guide the respondents to the right question. When the questionnaire uses the same conventions from beginning to end, it trains respondents in the use of those conventions. Unfortunately, it is sometimes necessary to depart from the conventions set earlier midway through a questionnaire. For example, the questions may switch response formats. According to the second guideline, in such cases, the visual cues should make it very clear what the respondents are supposed to do. Figure 7.4 makes the switch from circling the number of the selected answer to writing in the answer fairly obvious; the contrast between the white space for the response and the shaded background calls attention to what the respondents are now supposed to do and where they are supposed to do it.

The next two guidelines refer to the placement of instructions. The first page of a self-administered form often consists of lengthy instructions for completing the form. Jenkins and Dillman argue that respondents may simply skip over this initial page and start answering the questions; or, if they do read the instructions, the respondents may well have forgotten them by the time they actually need to apply them. Respondents are more likely to notice and follow instructions when

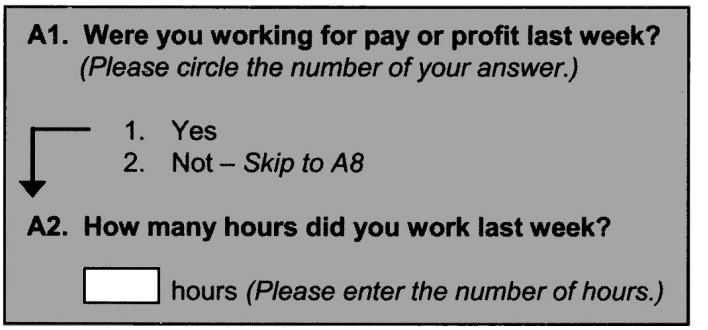

Figure 7.4 Illustration of use of visual contrast to highlight the response box.

they are placed right where they are needed. In Figure 7.4, instructions about how to indicate one's answer (*Please circle the number of your answer*) comes just before the answer options themselves. A related point is to put conceptually related information physically together. For example, a respondent should not have to look at the question and then look at the column headings to figure out what sort of answer is required. All of the information needed to understand the question should be in a single place.

Survey questions often try to cover multiple possibilities in a single item. For example, the question about doctor visits on page 228 is complicated partly because it covers doctors and other medical staff and face-to-face and telephone consultations all in a single question. The temptation to ask multiple questions in a single item can be especially strong in a self-administered questionnaire because adding an item or two can mean adding a page to the form. But asking multiple questions at once can impose a heavy interpretive burden on the respondents, who may be unable to keep the full set of logical requirements in mind. Here is an example discussed by Jenkins and Dillman:

> How many of your employees work full-time and receive health insurance benefits, how many work full-time without health insurance benefits, and what is the average length of time each type of employee has worked for this firm?

This item asks four separate questions: How many full-time employees get health insurance benefits? On average, how long have they worked for the firm? How many full-time employees do not get health insurance benefits? On average, how long have they worked for the firm? Whatever savings in space this achieves is likely to be offset by losses in understanding.

7.5 SUMMARY

Survey respondents engage in a series of processes involving comprehension, memory retrieval, judgment and estimation, and reporting in the course of

answering survey questions. Some visually presented self-administered questionnaires also require the respondent to make navigational decisions that can affect the flow of questions. Measurement of behaviors and attitudes seem to raise somewhat different issues at the judgment and estimation step.

The survey methodological literature contains many randomized experiments that demonstrate how measurement errors can arise at each of the steps in the response process. These include encoding problems, where the information sought is not stored in memory in an accessible form; misinterpreting the question because of wording or grammatical problems; forgetting and other memory problems; estimation issues in behavioral frequency questions; and judgment effects arising from question context or the sensitivity of the question. Over time, survey methodology has discovered tools to combat these problems. The tools vary according to whether the question is a nonsensitive behavior question, a sensitive behavior question, or an attitude question.

Alert readers will have noticed that this chapter spends about twice as much time cataloguing the problems that affect questions than in detailing their cures. There are several reasons for this, but the main one is that guidelines have their limitations. Accordingly, we think the principles from which the guidelines are derived are more important than the guidelines themselves.

What are these limitations of guidelines? First, any set of guidelines, no matter how comprehensive, cannot cover every situation. For example, some survey researchers would argue that certain types of questions—questions about causality or questions about our reactions to hypothetical situations—are basically too difficult to yield reliable answers. The guidelines we have presented omit this useful advice. It is impossible to formulate rules that cover every possible contingency. A second problem with guidelines is that every rule has its exceptions. Sudman and Bradburn recommend against check-all-that-apply items, and we generally agree with that advice, but the Census 2000 short form used that approach in collecting data on race. Most survey researchers would agree that this was a better solution than asking people if they are White (Yes or no?), Black (Yes or no?), and so on through the list. The check-all-that-apply format gave people a natural and efficient way to indicate a multiracial background. Still another limitation on guidelines is that they sometimes offer conflicting advice. On the one hand, we are supposed to specify the attitude object clearly but, on the other, we are supposed to avoid double-barreled items. Our guess is that many double-barreled items result from the effort to nail down the attitude object of interest. Similarly, the effort to spell out vague everyday concepts (like doctor visits) can lead to excessive complexity (see our discussion of the question on page 228). The issue of conflicts between guidelines is an important one. Often, such conflicts represent trade-offs between different, equally valid design considerations. For example, including a middle option in an attitude item has the advantage that it lets people who are actually in the middle accurately convey their views; it has the drawback that it provides an easy out for respondents who do not want to work out their position on the issue. It is not always easy to say which of the two considerations should take precedence. Or, to cite another issue, breaking a complicated item into simpler constituents may improve the answers but increase the length of the questionnaire.

Ultimately, guidelines are simply about what will work well in a given situ-

ation. These hypotheses should be tested whenever possible. Even the most experienced questionnaire designers like to have data to help them make decisions about questionnaires; after all, the proof of the pudding is in the eating. In the next chapter, we turn to methods for testing questionnaires.

KEYWORDS

acquiescence
aided recall
analogue method
bipolar approach
bounding
check-all-that-apply
closed questions
comprehension
double-barreled items
encoding
estimation
excessive complexity
false inference
faulty presupposition
generic memory
grammatical ambiguity
impression-based estimation
judgment
loading
navigational error
open questions
positivity bias
primacy effect
proxy
questionnaire
randomized response technique
ranking
rate-based estimation
recall-and-count
recency effect
reconstruction
reference period
rehearsal
reporting
retrieval
retrieval failure
retrieval cue
satisficing
sensitive question
social desirability
telescoping
unfamiliar term
vague concepts
vague quantifier

FOR MORE IN-DEPTH READING

Sudman, S., and Bradburn, N. (1982), *Asking Questions: A Practical Guide To Questionnaire Design*, San Francisco: Jossey-Bass.

Sudman, S., Bradburn, N., and Schwarz, N. (1996), *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*, San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000), *The Psychology of Survey Responses*, Cambridge: Cambridge University Press.

EXERCISES

1) In an effort to gauge public support for energy conservation, the (fictional) Andrews Foundation conducted a recent poll that found that 72% of Americans agreed with the following statement (phrased as an "Agree/Disagree" question):

> "I would support President Obama's decision to use the U.S. military to help local cities achieve energy independence by installing more energy efficient public lighting."

   a) What are the critical aspects of the attitude measured by this question?
   b) Does it meet the analytic goal of assessing public support for installing more energy efficient public lighting? Why or why not?

2) Using what you know about constructing attitude questions, write a standardized, interviewer administered question that you think will capture the direction and strength of public support for a ground invasion of Iraq involving U.S. troops. You may use more than one question, and any question and response format you desire. Be sure to specify response categories (if any are used), and what information is read to respondents (as opposed to interviewer instructions). Identify any skip patterns with appropriate formatting or notation.

   There are many different ways you could approach this task. To get started, you are free to do a search for questions used in actual surveys on this topic (e.g., you might explore the Gallup or Pew Research Center websites). If you use any items from other surveys; however, make sure that you identify (1) which items these are and where you took them from, (2) how, if at all, they have been revised, and (3) how your revisions meet the analytic goals of your questionnaire. Note: just because a question has been used and is in the public domain does not mean it is a good question. Your questionnaire should be based on the principles and rules of thumb you have learned in this chapter.

3) Write one or two paragraphs discussing how your question or questions are derived from the guidelines provided. Be specific. For example, if you used more than one question, why did you start with the question you did? If you used an 11-point response scale, why did you? How did you handle the tradeoffs between the different options available to you, and how might these affect the results of your questionnaire?

4) How does social desirability affect response? Describe two ways you could reduce the effects of social desirability when asking respondents to report their attendance of religious services.

5) Thinking about the cognitive processing models of the response task, describe potential problems you see in the wording and response options for this question:

Many people who own vehicles have regular service work done on them, such as having the oil changed. What kind of service work do you usually have done on your vehicle? Specify one or two. [Respondent reads the following response options to choose from.]

Oil changes
Fluid replacement
Tune-up
Body repair
Warranty-related services
Tire care
Transmission overhaul
Air conditioner treatment

6) For each of the items below, diagnose the problem(s) with the questions. Your diagnosis should be either the cognitive model of information processing presented in this chapter. Then, based on that diagnosis, suggest improved wording that solves the problem.

Ex. 1

During the past four weeks, beginning [DATE FOUR WEEKS AGO] and ending today, have you done any housework, including cleaning, cooking, yard work, and household repairs, but not including any activities carried out as part of your job?

Ex. 2

In the past week, how many times did you drink alcoholic beverages?

Ex. 3

Living where you do now and meeting the expenses you consider necessary, what would be the smallest income (before any deductions) you and your family would need to make ends meet each month?

Ex. 4

During the past 12 months, since [DATE], about how many days did illness or injury keep you in bed more than half the day? Include days while you were an overnight patient in a hospital.

7) Briefly give three reasons why it may be wise to avoid questions of the agree–disagree form.

8) Specify what type of estimation strategy the respondent might use in each of the following cases and its consequence on reported frequency:
a) Number of times the respondent was hospitalized in the past 2 years

b) Number of times the respondent ate in a restaurant in the past month
c) Number of time respondent's spouse/partner went on vacation during the past summer

9) Describe briefly two approaches the researcher can use to deal with telescoping.