Articles

Articles should deal with topics applicable to the broad field of program evaluation. Articles may focus on evaluation methods, theory, practice, or findings. In all cases, implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation, descriptions of a current evaluation study, critical reviews of some area of evaluation practice, and presentations of important new techniques. Manuscripts should follow APA format for references and style. Most submissions are 20–30 double-spaced typewritten pages in length; longer articles will also be published if their importance to AJE readers is judged to be high.

Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction

NORBERT SCHWARZ and DAPHNA OYSERMAN

ABSTRACT

Evaluation researchers frequently obtain self-reports of behaviors, asking program participants to report on process and outcome-relevant behaviors. Unfortunately, reporting on one's behavior poses a difficult cognitive task, and participants' reports can be profoundly influenced by question wording, format, and context. We review the steps involved in answering a question about one's behavior and highlight the underlying cognitive and communicative processes. We alert researchers to what can go wrong and provide theoretically grounded recommendations for pilot testing and questionnaire construction.

INTRODUCTION

Evaluation researchers make extensive use of self-reports of behavior at every phase of an evaluation project, including needs assessment, service utilization, program process, and

Norbert Schwarz • University of Michigan, Institute for Social Research, Ann Arbor, Michigan 48106-1248, USA; Tel.: (734) 647-3616; Fax: (734) 647-3652; E-mail: norbert.schwarz@umich.edu.

American Journal of Evaluation, Vol. 22, No. 2, 2001, pp. 127–160.
ISSN: 1098-2140

outcomes evaluation. For example, they may ask program participants to report on the number of cigarettes smoked, the amount of alcohol drunk, the frequency of fights with parents, the time spent doing homework, the frequency of service utilization, and a myriad of other behaviors. Although evaluators can obtain information about some behaviors from other sources, they typically must rely on self-reports to learn about many of the behaviors an intervention targets. In some cases, the cost of behavioral observation would be prohibitive and in others the behaviors are so infrequent, or so poorly observed by others, as to make anything but self-report impractical.

Unfortunately, a large body of research indicates that self-reports can be a highly unreliable source of data. Even apparently simple behavioral questions pose complex cognitive tasks, as our review will illustrate. Moreover, self-reports are highly context dependent and minor changes in question wording, format, or order can profoundly affect the obtained results. Hence, *how* evaluators ask a question can dramatically influence the answers they receive. Nevertheless, the psychology of asking and answering questions is largely absent from evaluation textbooks and rarely becomes a topic in the field's methodological literature. This dearth is probably a natural consequence of the expense of conducting evaluations—providing an intervention is resource intensive, as are scientifically sound samples and continuous tracking efforts. These constraints often force researchers to conduct evaluations at the edge of statistical power, making evaluators unwilling to experiment with the questionnaire format in their own studies and making them keen on using whatever comparison information is available from other studies, even when the questions used were less than optimal. Although this state of affairs is unlikely to change in the near future, there is a growing body of research outside of the evaluation domain that can help evaluators in designing better questionnaires.

Since the early 1980s, psychologists and survey methodologists have engaged in a collaborative research effort aimed at understanding the cognitive and communicative processes underlying question answering. Drawing on theories of language comprehension, memory, and judgment, they formulated models of the question answering process and tested these models in laboratory experiments and split-sample surveys (for comprehensive reviews see Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000; for research examples see the edited volumes by Hippler, Schwarz, & Sudman, 1987; Jabine, Straf, Tanur, & Tourangeau, 1984; Jobe & Loftus, 1991; Schwarz, Park, Knäuper, & Sudman, 1999; Schwarz & Sudman, 1992, 1994, 1996; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999; Tanur, 1992). This article reviews key lessons learned from this research, focusing on self-reports of behavior. To set the stage, we first contrast evaluators' hopes about the question-answering process with the reality experienced by participants attempting to answer these questions. Next, we review the key tasks involved in answering questions about one's behavior, identify the underlying processes, and discuss their implications for questionnaire construction. Where available, we highlight methods that are helpful at the question development and pilot testing stages, allowing evaluators to identify likely problems before they go into the field.

EVALUATORS' HOPES AND PARTICIPANTS' REALITY

Evaluators frequently ask questions such as, "Have you ever drunk beer, wine, wine coolers, whiskey, gin, or other liquor?" and "How many times have you had beer, wine, or other

TABLE 1.
Respondents' Tasks in Responding to a Question

Step 1: Understanding the question
Step 2: Recalling relevant behavior
Step 3: Inference and estimation
Step 4: Mapping the answer onto the response format
Step 5: "Editing" the answer for reasons of social desirability

liquor in the past month?" (adapted from Park, Kosterman, Hawkins, Haggerty, Duncan, Duncan, & Spoth, 2001). In posing such questions, researchers implicitly hope that participants will (1) understand the question, (2) identify the behavior of interest, and (3) retrieve relevant instances of the behavior from memory. When the question inquires about the actual frequency of the behavior, researchers further hope that participants (4) correctly identify the relevant reference period (e.g., "last month"), (5) search this reference period to retrieve all relevant instances of the behavior, (6) correctly date the recalled instances to determine whether they fall within the reference period, and (7) correctly add up all instances of the behavior to arrive at a frequency report. Once participants have determined the frequency of their behavior, they are (8) often required to map this frequency onto the response alternatives provided by the researcher. Finally, participants are expected to (9) candidly provide the result of their recall effort to the interviewer. Implicit in these—rarely articulated—hopes is the assumption that people *know* what they do and *can* report on their behavior with candor and accuracy, although they may not always be willing to do so. From this perspective, the evaluator's key task is to ask clear questions about meaningful behaviors in a setting that allows for candid reports.

Unfortunately, cognitive research suggests that respondents are rarely able to live up to the researchers' hopes. At the question comprehension stage, even apparently simple questions such as "What have you done today?" are highly ambiguous, as we shall see below. Moreover, recalling relevant behaviors from memory often takes considerable time, yet most research interviews allocate less than a minute to each question asked. More problematic, frequent behaviors are poorly represented in memory, and individual instances are difficult to retrieve, even with considerable time and effort, making a "recall-and-count" strategy unfeasible. Hence, respondents need to rely on a variety of estimation strategies to arrive at a meaningful estimate. Complicating things further, the response alternatives presented by the researcher may provide information that respondents use in interpreting the question asked and may suggest estimation strategies that systematically bias the obtained reports.

This article reviews these and related complications in some detail and highlights their implications for questionnaire construction. Its organization follows the sequence of participants' tasks, as shown in Table 1 (for variations on these themes, see Sudman et al., 1996; Tourangeau et al., 2000). Participants first have to understand the question to determine which behavior they are to report on (Step 1: Understanding the question). To do so, they draw on a wide range of contextual information in ways that researchers are often unaware of. Next, participants have to recall information about their behavior from memory (Step 2: Recalling relevant behavior). We discuss what participants can and cannot remember and review different strategies that researchers may employ to facilitate participants' recall. In most cases, however, recall will at best be fragmentary, and participants will need to apply various inference and estimation strategies to arrive at an answer (Step 3: Inference and

estimation). Having arrived at an answer in their own minds, participants can usually not report this answer in their own words. Instead, they need to map it onto the response alternatives provided by the researcher (Step 4: Mapping the answer onto the response format). Finally, participants may hesitate to candidly report their answer because of social desirability and self-presentation concerns and may hence "edit" their answer at this stage (Step 5: "Editing" the answer).

Two caveats are needed before we proceed. First, controlled experiments testing the effects of different question formats are rare in evaluation research. Accordingly, we draw on research examples from other domains to illustrate the basic cognitive and communicative processes underlying self-reports of behavior. Second, readers who hope for a list of simple "recipes" are likely to be disappointed. Although we provide recommendations throughout this article, these recommendations always need to be evaluated in the context of the specific research task at hand. Few recommendations hold under all conditions, and most involve tradeoffs that a researcher may or may not want to make. As is the case for any other research design decision, there is no alternative to thinking one's way through the complex issues at hand. Hopefully, our review of the basic cognitive and communicative processes involved in answering questions about one's behavior will provide readers with a useful framework for doing so.

STEP 1: UNDERSTANDING THE QUESTION

The key issue at the question comprehension stage is whether participants' interpretation of the question matches what the evaluator had in mind: Is the behavior that participants identify the one that the evaluator wanted them to report on? Even for simple and apparently straightforward questions, this is often not the case. For example, Belson (1981) observed that survey respondents' interpretation of "reading a magazine" covered a wide range of different behaviors, from having seen the magazine at a newsstand to having read it cover-to-cover or having subscribed to it. Given such variation in question comprehension, the question that a respondent answers may not be the question that the evaluator wanted to ask, nor do the answers provided by different respondents necessarily pertain to the same behavior. Moreover, divergent interpretations may result in underreporting (e.g., by respondents who read some articles but adopt a cover-to-cover interpretation) as well as overreporting (e.g., by respondents who adopt a saw-it-at-the-newsstand interpretation).

To avoid such problems, textbook discussions of questionnaire construction urge researchers to avoid unfamiliar and ambiguous terms (for good advice in this regard, see Sudman & Bradburn's *Asking Questions* [1983]). Although sound, this advice is insufficient. Even when all terms are thoroughly familiar, respondents may find it difficult to determine what they are to report on. Suppose, for example, that program participants are asked, "What have you done today?" Although they will certainly understand the words, they still need to determine what the researcher is interested in. Should they report, for example, that they took a shower or not? As this question illustrates, understanding the words, that is, the *literal meaning* of a question, is not sufficient to answer it. Instead, an appropriate answer requires an understanding of the *pragmatic meaning* of the question, that is, an understanding of the questioner's communicative intentions: What does the questioner want to know?

Participants infer what the questioner wants to know by bringing the tacit assumptions that underlie the conduct of conversations in everyday life to the research situation (for

reviews see Clark & Schober, 1992; Schober, 1999; Schwarz, 1996). These tacit assumptions have been explicated by Paul Grice (1975), a philosopher of language (for an introduction, see Levinson, 1983). His analysis shows that conversations proceed according to an overarching cooperativeness principle that can be described in the form of several maxims. A *maxim of relation* asks speakers to make their contribution relevant to the aims of the ongoing conversation. In daily life, we expect communicators to take contextual information into account and to draw on previous utterances in interpreting later ones. Yet, in standardized research situations this "normal" conversational behavior is undesired, and researchers often expect respondents to interpret each question in isolation. This, however, is not what respondents do, giving rise to context effects in question interpretation, as we shall see below. A *maxim of quantity* requests speakers to make their contribution as informative as is required, but not more informative than is required. This maxim invites respondents to provide information the questioner seems interested in, rather than other information that may come to mind. Moreover, it discourages the reiteration of information that has already been provided earlier, or that "goes without saying." A *maxim of manner* holds that a speaker's contribution should be clear rather than obscure, ambiguous, or wordy. In research situations, this maxim entails an "interpretability presumption." That is, research participants assume that the researcher "chose his wording so they can understand what he meant—and can do so quickly" (Clark & Schober, 1992, p. 27). Participants therefore assume that the most obvious meaning is likely to be the correct one, and if they cannot find an obvious meaning, they will look to the immediate context of the question to determine one.

The influence of such tacit maxims of conversational conduct is particularly pronounced in standardized research and evaluation settings. In daily life, we can ask a questioner for clarifications. But an interviewer who has been instructed not to violate the standardized script may merely reiterate the identical question, leaving it to the participant to make sense of it, and when participants face a self-administered questionnaire, nobody may be available to be asked. As a result, pragmatic inferences play a particularly prominent, but often overlooked, role in research settings as the following examples illustrate.

Pragmatic Inferences

To infer the intended meaning of a question, participants attend to a wide range of cues, of which we address the format and context of the question, the nature of the response alternatives, as well as information about the researchers' affiliation and the sponsor of the study (for more detailed reviews, see Schwarz, 1994, 1996).

Open versus closed question formats. With the above conversational maxims in mind, let us return to the question, "What have you done today?" Suppose that this question is part of an evaluation of a drop-in center for people with serious mental illness. The evaluator's goal is to assess whether the center helps structure participants' day and increases their performance of daily social and self-maintenance behaviors. To avoid cues that may increase socially desirable responding, the evaluator has deliberately chosen this open-ended global question. Which information are program participants and control respondents likely to provide?

Most likely, program participants will be aware that daily self-maintenance behaviors are of interest to the researcher and will consider their performance of these behaviors noteworthy, given that they are just reacquiring these routines. Hence, program participants

are likely to report these behaviors in response to the global question, "What have you done today?" In contrast, a control group of non-participants is unlikely to infer that the researcher is interested in "things that go without saying," such as taking a shower or brushing one's teeth, and may therefore not report these behaviors. As a result of these differential assumptions about what constitutes an "informative" answer, even a low level of self-maintenance behaviors among program participants may match or exceed the reports obtained from the control group, erroneously suggesting that the drop-in center is highly successful in helping its clients to return to normal routines. Similarly, drop-in participants who maintained daily self-maintenance behaviors may find them less noteworthy than participants who just reacquired these skills, raising additional comparison problems.

As an alternative approach, the evaluator may present a closed-ended list of daily self-maintenance behaviors. On the positive side, such a list would reduce the ambiguity of the open-ended question by indicating which behaviors are of interest to the researcher, ensuring that control respondents report on behaviors that otherwise "go without saying." On the negative side, the list would also provide program participants with relevant cues that may increase socially desirable responding. In addition, the list would remind both groups of behaviors that may otherwise be forgotten. As a result of these influences, any behavior is more likely to be endorsed when it is presented as part of a closed-ended question than when it needs to be volunteered in response to an open-ended question. At the same time, however, a closed-ended list reduces the likelihood that respondents report activities that are not represented on the list, even if the list offers a generic "other" response. What's not on the list is apparently of little interest to the researcher, and hence not reported. Accordingly, open- and closed-ended question formats reliably result in different reports (for reviews, see Schuman & Presser, 1981; Schwarz & Hippler, 1991).

Although these tradeoffs need to be considered in each specific case, a closed-ended format is often preferable, provided that the researcher can ensure that the list is reasonably complete. When evaluating service utilization, for example, a closed-response format that lists all available services and asks respondents to check off whether or not they used each service will ensure that respondents consider each possible service. Although this reduces the risk that a used service goes unreported, it also increases the risk that participants will overreport rarely used services. We return to the latter issue in the section on recall strategies.

Frequency scales. Suppose the evaluator of an anger management or social skills program asks participants how frequently they felt "really irritated" recently. To answer this question, respondents have to determine what the researcher means by "really irritated." Does this term refer to major or to minor annoyances? To identify the intended meaning of the question, respondents may consult the response alternatives the researcher provided. If the response alternatives present low frequency categories, for example, ranging from "less than once a year" to "more than once a month," they convey that the researcher has relatively rare events in mind. If so, respondents may conclude that the question refers to major annoyances, which are relatively rare, and not to minor irritations, which are likely to be more frequent. Conversely, a scale that presents high frequency response alternatives, such as "several times a day," may suggest that the researcher is mostly interested in minor irritations because major annoyances are unlikely to be so frequent.

To test this assumption, Schwarz, Strack, Müller, and Chassein (1988) asked respondents to describe a typical irritation after they had answered the frequency question. As expected, respondents who had received a high frequency scale reported less extreme

irritations than respondents who had received a low frequency scale. Thus, identically worded questions can acquire different meanings when accompanied by different frequency alternatives. As a result, respondents who are exposed to different scales report on substantively different behaviors (for more examples, see Schwarz, 1996).

Because response scales carry meaning, evaluators need to consider the implications of the response scale for the behavior in question: Does the scale convey information that is likely to influence respondents' interpretation of the question in unintended ways? Note also that it is problematic to compare reports of the "same" behavior when these reports were provided along different response scales. To the extent that the scale influenced question interpretation, respondents may, in fact, report on different behaviors. Hence, comparisons across samples and sites cannot be made with confidence if the questions were not asked in precisely the same way, including the nature of the response scale and whether the response was open- or close-ended.

Reference periods. Similar meaning shifts can arise from changes in the reference period. Suppose, for example, that an evaluator asks participants, in an open-ended format, how often they felt depressed, angry, and so on during a specified time period. Respondents again need to infer what type of anger or other emotion the researcher has in mind. When an anger question pertains to "last year," they may conclude that the researcher is interested in major annoyances because minor annoyances would probably be forgotten over such a long time period. Conversely, when the "same" question pertains to "last week," respondents may infer that the researcher is interested in minor annoyances, because major annoyances may not happen every week. Consistent with this assumption, Winkielman, Knäuper, and Schwarz (1998) observed that respondents reported on more intense anger when the question pertained to a reference period of "1 year" rather than "1 week."

Moreover, respondents reported a lower frequency of anger for the 1-year period than would be expected on the basis of their reports for a one-week period. Taken by itself, this observation may simply reflect that respondents forgot some distant anger episodes. Yet, the differential extremity of their examples indicates that forgetting is only part of the picture. Instead, respondents actually reported on differentially intense and frequent types of anger, and this meaning shift contributed to their differential frequency reports.

As this example illustrates, the same question may elicit reports about different behaviors and experiences depending on the reference period used. In principle, researchers can attenuate the influence of the reference period by providing an example of the behavior of interest. Although this helps to clarify the intended meaning, examples carry the same risk as incomplete lists in a closed-question format. Examples may inappropriately constrain the range of behaviors that respondents consider. It is, therefore, best to choose a reference period that is consistent with the intended meaning and to test respondents' interpretation at the questionnaire development stage by using the cognitive interviewing techniques we address below. Most important, however, evaluators need to be aware that answers to the same question are of limited comparability when the question pertains to reference periods of differential length.

Question context. Suppose an evaluator of a family-based intervention asks, "How often in the past year have you fought with your parents?" What is the evaluator asking about: physical fights, fights that result in punishments, squabbles over whose turn it is to do the dishes, "silent" disagreements? We have already shown that the frequency scale and the

reference period influence the way respondents interpret what the evaluator is asking. In addition, respondents are likely to use the context in which an evaluator asks a question to infer the appropriate meaning for ambiguous terms. When we asked teens how often they "fight" with their parents, we observed lower rates of "fighting" when this question followed questions about delinquency than when it preceded them (Oyserman, unpublished data). When queried, it turned out that teens understood the term "fight" to mean a physical altercation in which they hit their parents when the question was presented in the context of questions about stealing, gang fights, and so on, but not otherwise. To take what may be a more obvious example, a term such as "drugs" may be interpreted as referring to different substances in the context of questions about one's health and medical regime than in the context of questions about delinquency.

Contextual influences of this type are limited to questions that are substantively related; however, whether questions are substantively related may not always be obvious at first glance. To identify such influences at the questionnaire development stage, it is useful to present the question with and without the context to different pilot-test participants, asking them to paraphrase the question's meaning. In most cases, this is sufficient to identify systematic shifts in question meaning, and we return to these methods below.

Researcher's affiliation. Just as the preceding questions may provide unintended cues about the nature of a question, so can the researchers' academic affiliation or the sponsor of the survey. For example, Norenzayan and Schwarz (1999) asked respondents to explain the causes of a case of mass murder they read about. When the questionnaire was printed on the letterhead of an "Institute for Personality Research," respondents' explanations focused on personality variables. When the same questionnaire was printed on the letterhead of an "Institute for Social Research," respondents focused more on social determinants of homicide. Consistent with the conversational maxims discussed earlier, respondents tailored their explanations to meet the likely interests of the researcher in an effort to provide information that is relevant in the given context. Similar context effects may be expected for the interpretation of behavioral questions, although we are not aware of an empirical demonstration.

To the extent possible, evaluators may want to avoid drawing attention to affiliations that cue respondents to particular aspects of the study. Few researchers would imprint their questionnaire with the heading "Youth Delinquency Survey," yet when the neutrally labeled "Youth Survey" comes with a cover letter from the "Institute of Criminology" little may be gained by the neutral label.

Safeguarding Against Surprises

As the preceding examples illustrate, answering a question requires an understanding of its literal as well as its pragmatic meaning. Accordingly, the traditional textbook focus on using the "right words" needs to be complemented by close attention to the informational value of other question characteristics, which can serve as a basis for respondents' pragmatic inferences. Unfortunately, most comprehension problems are likely to be missed by traditional pilot test procedures, which usually involve a few interviews under field conditions to see if any problems emerge. Whereas such procedures are likely to identify overly ambiguous terms and complicated wordings, none of the above examples—from Belson's (1981) "reading magazines" to the influence of response alternatives or reference periods—would be

likely to show up as a comprehension problem in regular interviews. In all cases, respondents arrive at a subjectively meaningful interpretation and are hence unlikely to complain. If so, comprehension problems would only be identified when respondents' answers are odd enough to alert the researcher, which is often not the case. Nevertheless, the question that respondents answered may not be the one the researcher had in mind.

Cognitive pilot tests. Fortunately, these problems can be identified at an early stage. As a first step, we urge evaluators to look over their draft questionnaires, asking themselves: "What may my respondents conclude from the context of each question, the reference period, the response alternatives, and similar features? Is this what I want them to infer?" Next, evaluators may check each question for common problems. Many survey organizations have experts devoted to this task. Lessler and Forsyth (1996) offer an extensive checklist that alerts researchers to typical problems at the questionnaire design stage.

Once corrections have been made based on such a review, respondents' interpretation of questions can be explored in relatively inexpensive pilot tests with a small number of respondents drawn from the target population (including both program participants and control respondents, where applicable). For this purpose, evaluators can use a variety of cognitive interviewing procedures, which were designed to gain insight into respondents' thought processes (for reviews, see the contributions in Schwarz & Sudman, 1996, and chapter 2 of Sudman et al., 1996). These procedures range from asking respondents to paraphrase the question to the use of extensive probes and think-aloud protocols (see DeMaio & Rothgeb, 1996; Willis, Royston, & Bercini, 1991).

A particularly efficient approach to testing respondents' understanding of key concepts is the use of vignettes. In the case of Belson's (1981) example of "reading magazines," a researcher could present respondents with little cards that describe different instantiations of "reading," including behaviors that the researcher does not want to include (such as "Bob saw the magazine at a newsstand"). Respondents can then be asked to determine how the actors in the vignettes should answer different versions of the "reading" question, giving the researcher insight into respondents' interpretation of the range of "reading" covered by different question wordings. The wording that results in correct responses for most of the vignettes is the one best suited for the task. Once this wording is identified, one may further need to ensure that its meaning does not shift when the question is presented in the context of the questionnaire, which may require additional pilot tests.

In the case of self-administered questionnaires, the pilot testing needs to include the intended graphical lay-out of the questionnaire, which may raise its own set of complications (for helpful advice, see Jenkins & Dillman, 1997). The importance of the pilot testing of graphical design is perhaps best illustrated by the 2000 Presidential elections in the United States. Had the ballot format used in Florida's Palm Beach County been properly tested, the closely contested race in Florida would probably have come out otherwise (see Sinclair, Mark, Moore, Lavis, & Soldat, 2000).

Most well-designed and extensively pilot-tested questions will nevertheless be interpreted in a different way by some respondents. Such idiosyncratic variation is unavoidable. But unintended systematic influences can be eliminated through appropriate cognitive pilot testing at relatively low cost.

Interviewing procedures. As noted earlier, strictly standardized interviewing procedures contribute to respondents' reliance on contextual information by discouraging the

interviewer from providing additional clarifications. Hence, respondents are left to their own devices and have little choice but to refer to the context to make sense of a question they do not understand. Recent experimental research demonstrates that many comprehension problems can be attenuated when interviewers are allowed to provide explanations, either when respondents ask for them or when the interviewer notices that the respondent may not have understood the intended meaning (see Schober, 1999; Schober & Conrad, 1997). Obviously, it is important in this case that the interviewer understands the question as intended, which cannot be taken for granted and requires appropriate interviewer training.

Note that this is not a recommendation to assess factual information through semistructured or qualitative interviews. Letting interviewers ask questions in any way they want does not guarantee comprehension. To the contrary, such procedures preclude the benefits of cognitive pilot testing and do not allow the researcher to optimize question wording. Moreover, procedures that result in different question orders for different respondents may introduce differential context effects in a way that is impossible to tract. Hence, we instead recommend the use of properly pilot-tested standardized wordings, in combination with interviewer instructions that allow the interviewer to provide additional clarifications when needed. Schober (1999) provides an informative discussion of these issues.

STEP 2: RECALLING RELEVANT BEHAVIOR

Once respondents understand what they are to report on, they need to retrieve relevant information from memory. In this section, we first review key lessons learned from autobiographical memory research, a field of psychology that addresses how people encode, store, and retrieve information about their own lives (for a readable introduction see, Conway, 1990). Subsequently we review what researchers can do to facilitate respondents' recall task.

Autobiographical Memory

In evaluation research, many questions about respondents' behavior are frequency questions, pertaining, for example, to how often they used a service or engaged in some risky behavior, such as drinking alcohol, driving without a safety belt, and so on. As already noted, researchers typically hope that respondents will identify the behavior of interest, scan the reference period, retrieve all instances that match the target behavior, and finally count these instances to determine the overall frequency of the behavior. However, respondents are unlikely to follow such a "recall and count" strategy, unless the events in question are highly memorable and their number is small (for a discussion, see Brown, in press). In fact, several factors render this strategy unsuitable for most of the behaviors in which evaluators are interested.

First, memory decreases over time, even when the event is relatively important and distinctive. For example, Cannell, Fisher, and Bakker (1965) observed that only 3% of their respondents failed to report an episode of hospitalization when interviewed within 10 weeks of the event, yet a full 42% did so when interviewed 1 year after the event. There may be little to be recalled once time has passed.

Second, when the question pertains to a frequent behavior, respondents are unlikely to have detailed representations of numerous individual episodes of a behavior stored in

memory. Instead, the various instances of closely related behaviors blend into one global, knowledge-like representation that lacks specific time or location markers (see Linton, 1982; Neisser, 1986; Strube, 1987). As a result, individual episodes of frequent behaviors become indistinguishable and irretrievable. This is most likely to occur for mundane behaviors of high frequency, but has also been observed for more important experiences. Mathiowetz and Duncan (1988), for example, found that respondents were more accurate in recalling a single spell of unemployment than they were at recalling multiple spells of unemployment. Throughout, the available research suggests that the recall of individual behavioral episodes is largely limited to rare and unique behaviors of considerable importance (see Conway, 1990; Strube, 1987).

Third, our autobiographical knowledge is not organized by categories of behavior such as "drinking alcohol" or the like. Instead, the structure of autobiographical memory can be thought of as a hierarchical network that includes *extended periods* (such as "the years I lived in New York") at the highest level of the hierarchy. Nested within these high-order periods are lower-level extended events pertaining to this time, such as "my first job" or "the time I was married to Lucy." Further down the hierarchy are *summarized events,* which correspond to the knowledge-like representations of repeated behaviors noted above (e.g., "During that time, my spouse and I quarreled a lot."). *Specific events,* such as a particular instantiation of a disagreement, are represented at the lowest level of the hierarchy. To be represented at this level of specificity, however, the event has to be rather unusual. As these examples illustrate, autobiographical memory is primarily organized by time ("the years in New York") and relatively global themes ("first job," "first marriage") in a hierarchical network (for a comprehensive review, see Belli, 1998). This network "permits the retrieval of past events through multiple pathways that work top-down in the hierarchy, sequentially within life themes that unify extended events, and in parallel across life themes that involve contemporaneous and sequential events" (Belli, 1998, p. 383). Thus, thinking of the "years in New York" would lead to information about the first job and first marriage (top-down) and thinking about the first marriage may prompt memories of a later marriage (within theme). Any specific event that comes to mind along the way may prompt memories of other events. Such searches take considerable time, and their outcome is somewhat haphazard, depending on the entry point into the network at which the search started. Hence, using multiple entry points and forming connections across different periods and themes improves recall.

Unfortunately, many of the behaviors researchers are interested in do not constitute meaningful themes that map onto individuals' autobiographical memory. Moreover, the format of most behavioral questions does not encourage extensive searches through the hierarchical network of autobiographical memory. One exception to this generalization is a method known as *Event History Calendars,* which we review in a later section.

Facilitating Recall

These basic aspects of autobiographical memory bear on how researchers can facilitate respondents' recall of relevant behaviors. We now review the most common procedures and note their promises and shortcomings.

Reference periods and recall cues. In general, memory decreases over time, rendering it difficult to recall distant events, unless they were highly important, unusual, and memorable. Recall improves, however, when helpful recall cues are available (Baddeley,

1990). Hence, researchers can, in principle, improve the likelihood of accurate recall by restricting the recall task to a short and recent reference period and by providing appropriate recall cues. But the emphasis is on "in principle," and there are important drawbacks to these strategies.

On the positive side, short reference periods make it more likely that respondents will try to recall relevant episodes, whereas long reference periods, which include a larger number of episodes, encourage guessing and estimation (e.g., Brown, in press; Blair & Burton, 1987). On the negative side, short reference periods may result in many "zero" answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with a high behavioral frequency. To counter this problem, one can increase the sample size to ensure that a sufficient number of low-frequency respondents had a relevant episode during the short reference period. However, this solution may come at considerable expense, or in the case of much evaluation research, be impossible given limits on sample size.

Similarly, appropriately selected cues typically improve respondents' recall. In general, the date of an event is the poorest cue, whereas cues pertaining to what happened, where it happened, and who was involved are more effective (e.g., Wagenaar, 1986, 1988). Yet, recall cues share many of the characteristics of closed-response formats and can constrain the inferred question meaning. Respondents may therefore limit their memory search to behaviors that are closely related to the recall cues and may omit behaviors that provide a poor match. It is therefore important to ensure that the recall cues are relatively exhaustive and compatible with the intended interpretation of the question.

Adding to these complexities, the length of the reference period and the specificity of the recall cues influences the number of episodes respondents are to report on: The more recall cues specify a particular behavior and the shorter the reference period is, the smaller is the number of relevant behavioral episodes. A small number of episodes, however, is likely to be systematically overestimated, whereas a large number of episodes is likely to be underestimated, as illustrated in the next section.

Decomposition strategies: Recall cues and estimation strategies. Closely related to the provision of recall cues is the decomposition of a complex task into several more specific ones. For example, a researcher may decompose a question about "drinking alcohol" into three questions about "drinking wine," "drinking beer," and "drinking liquor" or may even decompose "liquor" into additional subcategories. These decomposed questions provide the specific drinks as recall cues, reminding respondents of episodes that may otherwise be forgotten. Empirically, decomposed questions do, indeed, result in reliable increases in reported frequency (e.g., Blair & Burton, 1997; Sudman & Schwarz, 1989). That is, the sum across beer, wine, and liquor will be higher than the frequencies reported in response to the global alcohol question. Because researchers assume that forgetting is the key problem in retrospective reports, this observation seems to confirm the expected benefits of providing recall cues through more specific questions. Unfortunately, the available data do not provide strong support for this optimistic conclusion.

Although decomposition reliably increases the reported frequency of behaviors, it does not reliably increase the accuracy of the obtained reports (e.g., Belli, Schwarz, Singer, & Talarico, 2000). Instead, the increase in reported frequencies may often reflect a change in the underlying estimation processes. In general, people tend to overestimate the occurrence of low frequency events and to underestimate the occurrence of high-frequency events (see Fiedler & Armbruster, 1994), a variant of the "response contraction bias" observed in many

psychophysical studies (for a review, see Poulton, 1989). Because global questions (e.g., "drinking alcohol") pertain to more frequent behaviors than do more specific questions (e.g., "drinking liquor"), global questions foster underestimates. In contrast, a series of more specific and narrow questions ("drinking wine," "drinking beer," "drinking liquor") fosters small overestimates for each of the subcategories. This results in sizeable overestimates once the answers to the specific questions are added up to arrive at "drinking alcohol."

Consistent with this interpretation, any decomposition of a larger category into smaller subcategories results in increased behavioral reports, even when the decomposition creates subcategories that are known to be poor retrieval cues, such as dates or time of day. For example, Belli et al. (2000) compared participants' estimates of how many times they used the phone (a common behavior) with phone records. Consistent with many earlier findings, respondents reported fewer calls when the question pertained to "last week" than when they had to report separately on each day of the week; similarly, they reported fewer calls when the question pertained to "yesterday" than when yesterday was broken down into eight time periods. More important, however, record checks indicated that the decomposition of the general question did not increase the accuracy of respondents' recall; it only increased their frequency estimates.

At the present stage of research on decomposition, we conjecture that decomposing a general question into several more specific ones is useful when the specific questions pertain to infrequent and memorable behaviors. In this case, specific questions may provide helpful recall cues that may increase the accuracy of reports. When the specific questions pertain to frequent and mundane behaviors, however, little may be gained. These behaviors are poorly represented in memory and difficult to retrieve under any circumstances. Hence, respondents have to rely on estimation strategies. In doing so, they are likely to overestimate the frequency of each specific behavioral category, resulting in pronounced overestimates once specific categories are added up to arrive at an estimate for the general class of related behaviors.

Time and motivation. In general, recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds (e.g., Reiser, Black, & Abelson, 1985), and repeated attempts to recall may result in the retrieval of additional material, even after a considerable number of previous trials (e.g., Williams & Hollan, 1981). Unfortunately, respondents are unlikely to have sufficient time to engage in repeated retrieval attempts in most research situations. Moreover, they may often not be motivated to do so even if they had the time.

Accordingly, explicitly instructing respondents that the next question is really important, and that they should do their best and take all the time they may need, has been found to improve recall (e.g., Cannell, Miller, & Oksenberg, 1981). Such instructions are particularly important in telephone interviews, where respondents (and interviewers) are uncomfortable with the moments of silence that accompany extended memory search. One way to provide respondents with more time is the use of redundant questions. Instead of asking, "Which of the following services have you used during the last month?" one might ask, "Our next question is about the services you have used during the last month. I will read you a list of services. We'd like you to tell us for each one if you have used it during the last month." This strategy has been found to improve the accuracy of recall to some extent, as have been explicit encouragement to take all the time needed (for a review, see Cannell et al., 1981).

Instructing respondents that the task is important and that they should take their time to

arrive at an accurate answer is one of the most efficient low-cost strategies a researcher can employ. Note, however, that it needs to be employed sparingly and may lose its credibility when used for too many questions within an interview.

Temporal direction of search. Less intuitively obvious is that the direction in which respondents search memory may influence the quality of recall. Specifically, better recall is achieved when respondents begin with the most recent occurrence of a behavior and work backward in time than when they begin at the beginning of the reference period working forward in time (e.g., Loftus & Fathi, 1985; Whitten & Leonard, 1981). This presumably occurs because memory for recent occurrences is richer, and the recalled instances may serve as cues for recalling previous ones. This advantage may not be observed, however, when the material has an inherent temporal or causal order, that is, where preceding events logically lead to subsequent ones. In the latter case, a chronological order of recall may be preferable, although relevant experimental comparisons are not available.

Dating recalled instances. Suppose, optimistically, that respondents have successfully recalled or reconstructed several specific instances of the behavior under study. To move from these instances to a frequency report, they need to determine for each instance whether it occurred during the reference period. This requires that they understand the boundaries of the reference period and that they can accurately date each instance relative to that period.

Reference periods that are defined in terms of several weeks or months are highly susceptible to misinterpretations. For example, the term "during the last 12 months" may be interpreted as a reference to the last calendar year, as including or excluding the current month, and so on (Bradburn et al., 1987). Similarly, anchoring the reference period with a specific date, for example, "Since March 1, how often. . . ?," is not very helpful because respondents will usually not be able to relate an abstract date to meaningful memories.

A more efficient way to anchor a reference period is the use of salient personal or public events, often referred to as "temporal landmarks" (Loftus & Marburger, 1983). Unfortunately, meaningful landmarks are not always available, although prominent holidays (such as New Year's or Labor Day) can often be used. In that case, it is helpful to ask respondents what they have done during that holiday, to evoke a specific memory that may serve as an anchor for the reference period.

Even under optimal conditions, however, event dating is likely to reflect both "forward" and "backward telescoping." That is, distant events are assumed to have happened more recently than they did (called "forward telescoping"), whereas recent events are assumed to be more distant than they are (called "backward telescoping," for reviews and a theoretical model, see Bradburn, Huttenlocher, & Hedges, 1994; Sudman et al., 1996, chapter 8).

Combining Helpful Strategies: The Event History Calendar

Although the above strategies improve recall to some extent, they fail to take full advantage of what has been learned about the hierarchical structure of autobiographical memory. A promising alternative approach is offered by the *event history calendar* (for a comprehensive review, see Belli, 1998). This method is also known as the *life history calendar* (Axinn, Pearce, & Ghimire, 1999; Caspi, Moffitt, Thornton et al., 1996; Freedman, Thornton, Camburn, Alwin, & Young-DeMarco, 1988) or *life chart interview* (Lyketsos,

Nestadt, Cwi, Heithoff, & Eaton, 1994). It allows respondents to place their behavior in time and space and uses the hierarchically nested structure of autobiographical memory to facilitate recall. Moreover, it provides respondents with considerable time for the recall task and emphasizes the importance of accuracy. Finally, it explicitly encourages the correction of earlier answers as newly recalled information qualifies earlier responses. This correction opportunity is missed under regular interview formats, where respondents rarely return to earlier questions.

Initially developed to assess extended periods of life, event history calendars can be adapted to any time period. To help respondents recall their alcohol consumption during the last week, for example, respondents may be given a calendar grid that provides a column for each day of the week, cross-cut by rows that pertain to relevant contexts. For example, they may be asked to enter for each day what they did, who they were with, if they ate out, and so on. Reconstructing the last week in this way provides a rich set of contextual cues, with entries in one row often prompting memories relevant to a different row. Based on this rich network of associations, individual episodes are more likely to be retrieved than under any other method, and any given episode may prompt additional memories.

However, to date, most applications of event history calendars have focused on more extended time frames, such as respondents' life-, employment-, or health-history. In this case, respondents may begin by marking life periods, such as being in school, living at home, getting a first job, and so on. Next, they may be asked to mark other events within these periods, changing the timing of already marked events as needed when newly recalled information requires corrections. Within the developing rich structure of associations, respondents are usually able to recall and date events with considerable accuracy. For example, Freedman et al. (1998; see also Caspi et al., 1996) observed high accuracy in the reconstruction of past life periods (such as attending school, living with one's parents, first job, etc.) when they compared respondents' reports to previously collected concurrent data as part of a longitudinal multigenerational study.

Although the usefulness of event history calendars has been primarily demonstrated for the long-term recall of major events (such as employment histories, criminal histories or illness histories), the method can be adapted to shorter time periods and the assessment of more mundane behaviors (for a review, see Belli, 1998). Although costly in terms of interview time, we consider such adaptations to be among the most promising developments in the assessment of behavioral reports.

Safeguarding Against Surprises

Many recall questions would never be asked if researchers first tried to answer them themselves. Answering the questions one intends to ask is, therefore, an important first step. If you find it difficult, despite all the motivation you bring to the task, your respondents will probably find it next to impossible. Nevertheless, they will play by the rules and provide an answer. But little is gained if this answer is error ridden and bears little resemblance to reality. It is, therefore, better to lower one's goals and to design a more realistic, limited, and less demanding recall task than to pursue an ideal data set that exceeds respondents' abilities.

To explore what respondents can and cannot report evaluators can draw on cognitive interviewing techniques to identify likely recall and estimation problems in pilot tests (for reviews, see Schwarz & Sudman, 1996; Sudman et al., 1996, chapter 2; Willis et al., 1991). Most promising is the use of think-aloud protocols that can provide insight into respondents'

recall and reconstruction processes. Alternatively, respondents can be asked to describe how they arrived at an answer. In any case, pilot test respondents should be invited to comment on the difficulty of the task and the confidence they have in their answer. More often than not, the experience will be a sobering one, forcing researchers to adjust their (usually unrealistic) expectations.

Unfortunately, there is no "silver bullet" to solve recall problems, and most strategies employed to improve recall come with their own tradeoffs, as highlighted above. In all situations, however, researchers are well advised to instruct respondents that accurate recall is important and to encourage them to take their time. Moreover, cognitive pilot testing can help in determining suitable reference periods, which are usually shorter than researchers expect. Keep in mind, however, that the reference period may influence question interpretation, as discussed earlier. In addition, question sequences that bring to mind the various contexts in which the behavior is likely to be engaged in are helpful; the format of event history calendars provides an excellent solution for this issue. Finally, respondents should be encouraged to correct earlier answers once pertinent information comes to mind later on; again, the format of event history calendars facilitates this.

No matter how much effort we put into question design, however, the best we can usually hope for is a reasonable estimate, unless the behavior is rare and of considerable importance to respondents. Next, we turn to respondents' estimation strategies.

STEP 3: INFERENCE AND ESTIMATION

Given the reviewed difficulties of recalling information about one's behavior from memory, it is not surprising that respondents usually resort to a variety of inference strategies to arrive at a plausible estimate (Conrad & Brown, 1996; Sudman et al., 1996, chapter 9). Even when respondents can recall relevant episodic information, the recalled material may not cover the entire reference period, or respondents may be aware that their recall is likely to be incomplete. In such cases, they may base their inferences on the recalled fragments, following a strategy that is often referred to as "decomposition." In other cases, respondents may draw on subjective theories that bear on the behavior in question. When asked about past behavior, for example, they may ask themselves if there is reason to assume that their past behavior was different from their present behavior, if not, they may report their present behavior as an approximation. Similarly, when asked about the behavior of others, they may draw on their impression of "what kind of person" the other is, basing the estimate on an implicit theory of personality. Finally, respondents may extract relevant information from the questionnaire, for example, by using the response alternatives as a frame of reference in thinking about their own behavior. We review these different strategies below.

Empirically, respondents' own uncertainty about the result of their estimation efforts often finds its expression in *rounded numbers.* In the case of frequency reports, the answers typically show heaps at multiples of 5 and 10 (for a discussion, see Tourangeau et al., 2000). Similarly, reports of elapsed time (e.g., "How many days ago. . . ?") usually show heaps of responses at prototypical values, such as 7 or 30, when the metric pertains to days (e.g., Huttenlocher, Hedges, & Bradburn, 1990). Because reality rarely comes in multiples of round numbers, distributions that peak at such numbers are good indications that respondents relied on estimation strategies and selected a "good enough" estimate. In fact, using such multiples may be a way of expressing that the answer is not an exact one (see Tourangeau et al., 2000).

Inferences Based on Partial Recall: Decomposition and Extrapolation

Many recall problems become easier when respondents break down, or "*decompose*," the recall task into several subtasks. The decomposition strategy that respondents are most likely to use spontaneously is the temporal decomposition of a long reference period into several smaller ones (for a review, see Sudman et al., 1996). To answer the question how many times she has had beer, wine, or alcohol in the last month, for example, a respondent may determine that she drinks about every weekend night, that is, unless no one she knows is having a party, which happened last weekend but, she thinks, not the week before. Thus, she may infer, this makes eight times that she usually drinks a month, but probably only six times for this month. Then she has to decide whether the researcher means to count each time she drinks or how many drinks; if the latter, then she will again estimate based on her recall of how much she usually drinks, resulting in a final count of say "18 times during the last month." Estimates of this type are likely to be accurate if the respondent's inference rule is adequate and if exceptions to the usual behavior are rare.

In fact, some behaviors are regular enough that respondents know their rate of occurrence, like "every Sunday" for attending church or "daily" for washing one's hair (see Brown, in press; Menon, 1993, 1994). In such cases, respondents can apply the rate to the time period and extrapolate to arrive at a correct answer. In the absence of such fortunate conditions, however, temporal decomposition-and-extrapolation strategies are unlikely to result in accurate estimates.

Moreover, a large body of research indicates that people overestimate the frequency of rare behaviors and underestimate the frequency of frequent behaviors (see Sudman et al., 1996). Given that any behavior is less frequent during a short time period than during a long time period, the reference period used is likely to affect the estimate in specific ways: Whereas a long reference period, such as "1 month," fosters underestimates, short reference periods foster overestimates, as discussed in the section on reference periods and recall cues.

Inferences Based on Subjective Theories

In the absence of relevant episodic information, respondents may draw on their general assumptions about the world to arrive at a plausible estimate. Psychologists often refer to such assumptions as "subjective theories." Here we address two that are of particular relevance, namely assumptions about stability and change in one's behavior and assumptions about another's personality.

Theories of stability and change: What my behavior must have been. To answer questions about past behaviors, respondents often use their current behavior as a benchmark and ask themselves if there is reason to believe that their past behavior was similar to, or different from, their present behavior. If they see no reason to assume their behavior has changed over time, they use their present behavior as an estimate of their past behavior. If they do believe their behavior has changed, they adjust the initial estimate based on their current behavior to reflect the assumed change. Our preceding discussion of decomposition has already illustrated these kinds of inferences. Not surprisingly, the resulting reports of past behavior are correct to the extent that respondents' subjective theories of stability and change are correct. Unfortunately, this is rarely the case (for a comprehensive review, see Ross, 1989).

In many domains, individuals assume an unrealistically high degree of stability, resulting in underestimates of the degree of change that has occurred over time. Accordingly, retrospective estimates of income (Withey, 1954) and of tobacco, marijuana, and alcohol consumption (Collins, Graham, Hansen, & Johnson, 1985) were found to be heavily influenced by respondents' income or consumption habits at the time of interview. On the other hand, when respondents have reason to believe in change, they will detect change, even though none has occurred (see Ross, 1989). For example, Ross and Conway (1986) had students participate in a study skills training that did not improve their skills on any objective measure (and was not expected to do so). Following the training, researchers asked participants to recall how skilled they were before the training. Applying a plausible theory of change, namely that the training improved their skills, participants inferred that their prior skills must have been much worse than they were after training. Hence, they retrospectively reported having had poorer pre-training skills than they indicated before the training, apparently confirming the intervention's success. This result was obtained despite incentives to respondents to recall their earlier answers as accurately as possible. As Ross and Conway (1986) noted, you can always get what you want by revising what you had.

This possibility is particularly troublesome for evaluation research, given that most interventions are likely to evoke a subjective theory of change. As a result, respondents may reconstruct their earlier behaviors as having been more problematic than they were, apparently confirming the intervention's success—provided they believe the intervention was likely to help them (a belief that entails a subjective theory of change). Conversely, they may reconstruct their earlier behaviors as having been less problematic, and closer to their current behaviors, if they believe the intervention was unlikely to help them (a belief that entails a subjective theory of stability). This issue deserves systematic investigation in the context of evaluation studies.

The recommendation emerging from this research will not come as a surprise to evaluators: Asking program participants to report on how their behavior has changed over the course of the intervention, or what their behavior was prior to the intervention, is likely to result in theory-driven reconstructions. These reconstructions are useless as measures of objective change, although they may be of interest as measures of participants' subjective perceptions. To assess actual change, we need to rely on before-after, or treatment-control, comparisons, and if we have missed asking the right question before the intervention, little can be done after the fact to make up for the oversight.

Theories of personality: Reporting on the behavior of others. Subjective theories play an even more prominent role in the inference process when respondents are asked to report on the behavior of others, in which case they rely on their general theories about "what kind of person" the other is. Reports about others' behaviors, often referred to as proxy-reports, may be sought because the target person is not available for an interview or because the researcher wants to validate a respondent's reports against the perceptions of a familiar other, often another household member. Researchers sometimes assume that these proxy-reports are more accurate than self-reports. However, controlled experimental studies provide little support for this conclusion (for a discussion, see Moore, 1988; Schwarz & Wellens, 1997). Instead, the problems associated with retrospective self-reports are compounded when respondents are asked to report about the behavior of others. In many cases, respondents may not be fully aware of the others' behavior. Moreover, others' behaviors are even more poorly represented in memory, unless they were extreme and memorable.

Empirically, studies based on the collection of self- and proxy-reports from two members of the same household obtained moderate degrees of agreement between self- and proxy-reports of daily behaviors (e.g., Mingay, Shevell, Bradburn, & Ramirez, 1994; Skowronski, Betz, Thompson, Walker, & Shannon, 1994; Sudmanet al., 1994). Not surprisingly, the agreement between self- and proxy-reports is highest for behaviors in which both household members participated, in which case proxy-respondents can draw on their memory for their own behavior to arrive at a report about the other household member. In these cases, the agreement between self- and proxy-respondents may reach $r = 0.8$. Agreement is lowest for behaviors that individuals performed in the absence of the proxy-respondent and were unlikely to discuss with the proxy (with *r*s hovering around 0.4). Behaviors that individuals performed without the proxy, but were likely to discuss frequently with the proxy, fall in between these extremes (with *r*s around 0.6). Although these correlations may seem quite comforting, it is important to realize that agreement between self- and proxy-reports may reflect reliance on similar estimation strategies, rather than accurate recall.

Note that self-reports are reports of an actor about his or her own behavior, whereas proxy-reports are reports of an observer about a well-known other's behavior. Accordingly, we may bring basic research on actor-observer differences in social perception (Jones & Nisbett, 1971) to bear on these tasks. As experimental research in social psychology demonstrated (for a review, see Watson, 1982), observers are more likely to draw on what they know about an actor's character or dispositions in explaining his or her behavior than is the actor him or herself. Hence, proxies derive their reports to a larger degree from their assumptions about the "kind of person" the actor is, a tendency that is compounded by their lack of situational knowledge when they did not themselves participate in the respective behavior. Consistent with this assumption, several experiments indicate that proxy-reports are more likely to be derived from dispositional information than are self-reports, which are more likely to be based on episodic information (Schwarz & Wellens, 1997). This difference between self- and proxy-respondents' strategies has important methodological implications.

First, given that proxy-respondents derive their answers from general knowledge about the actor (the "kind of person" he or she is), they arrive at similar answers to related questions. As a result, proxy-reports show higher internal consistency than do self-reports, apparently suggesting that proxy-reports are more reliable. Yet, this internal consistency merely reflects the underlying inference strategy. It should, therefore, not be taken as evidence for higher accuracy.

Second, drawing on the "kind of person" the actor is, proxy-respondents underestimate the variability of the actor's behavior over time. Accordingly, proxy-reports and self-reports of behavioral frequencies show low convergence for short and recent reference periods, for which the actor can draw on some episodic information in providing a self-report. As the actor's access to episodic information decreases because of longer or more distant reference periods, however, the actor has to rely on dispositional information as well. As a result, the convergence of self- and proxy-reports increases for long and distant reference periods (Schwarz & Wellens, 1997). Again, this increase merely reflects reliance on the same inference strategies and should not be taken as evidence for higher accuracy

In sum, in arriving at an answer proxy-respondents are likely to draw on their theories about the "kind of person" the actor is. They, therefore, underestimate the variability of the actor's behavior over time and situations, resulting in reports that have high internal consistency. Unfortunately, this consistency is not an indication of accuracy. Hence, proxy-reports inform us more about the proxy's impression of the actor rather than about the actor's

Low Frequency Scale		High Frequency Scale	
( )	*never*	( )	*twice a month or less*
( )	*about once a year*	( )	*once a week*
( )	*about twice a year*	( )	*twice a week*
( )	*twice a month*	( )	*daily*
( )	*more than twice a month*	( )	*several times a day*

Figure 1. Frequency response alternatives for reporting physical symptoms.

actual behavior. Accordingly, self-reports are clearly preferable. An exception to this generalization are sensitive behaviors that the actor may be unlikely to report for reasons of social desirability. In this case, the proxy-respondent's impression may still be more informative than the actor's guarded response, although it should be considered an informed impression rather than a factual report.

Inferences Based on the Research Instrument: Frequency Scales

In many studies, researchers ask respondents to report the frequency of their behavior by checking the appropriate alternative from a list of quantitative response alternatives of the type shown in Fig. 1. What is often overlooked is that participants assume that the researcher constructed a meaningful scale that is relevant to their task. Specifically, they assume that the scale reflects the researcher's knowledge about the distribution of the behavior, with values in the middle range of the scale corresponding to the "usual" or "average" behavior, and values at the extremes of the scale corresponding to the extremes of the distribution. Hence, the frequency values presented by the researcher influence respondents' own frequency estimates and subsequent judgments, as well as respondents' interpretation of the question (as discussed in the section on question comprehension).

Frequency estimates. Given the above assumptions, respondents can use the range of the response alternatives as a frame of reference in estimating their own behavioral frequency. When respondents use this strategy, they make higher frequency estimates if they are given a scale that presents high rather than low frequency response alternatives (See Fig. 1). For example, Schwarz and Scheuring (1992) asked 60 patients of a German mental health clinic to report the frequency of 17 symptoms along one of the two scales shown in Fig. 1. Across 17 symptoms, 62% of the respondents reported average frequencies of more than twice a month when presented with the high frequency scale, whereas only 39% did so when presented with the low frequency scale, resulting in a mean difference of 23 percentage points. The impact of response alternatives was strongest for the ill-defined symptom of "responsiveness to changes in the weather," where 75% of the patients reported a frequency of more than twice a month along the high frequency scale, whereas only 21% did so along the low frequency scale. Conversely, the influence of response alternatives was least pronounced for the better defined symptom "excessive perspiration," with 50% versus 42% of the respondents reporting a frequency of more than twice a month in the high and low frequency scale conditions, respectively.

This influence of frequency scales has been observed across a wide range of different behaviors, including health behaviors (e.g., Gaskell, O'Muircheartaigh, & Wright, 1994), television consumption (e.g., Schwarz, Hippler, Deutsch, & Strack, 1985), sexual behaviors

(e.g., Schwarz & Scheuring, 1988; Tourangeau & Smith, 1996), and consumer behaviors (e.g., Menon, Rhagubir, & Schwarz, 1995). For example, in a representative sample of American adults, Tourangeau and Smith (1996) observed that men and women reported more sexual partners when asked to report the number of sexual partners on a high, rather than a low, frequency scale.

As expected on theoretical grounds, the impact of response alternatives is more pronounced the more poorly the behavior is represented in memory, which forces respondents to rely on an estimation strategy. When the behavior is rare and important, and hence well represented in memory, the impact of response alternatives is small because no estimation is required. Finally, when a respondent engages in the behavior with high regularity (e.g., "every Sunday"), its frequency can easily be derived from this rate information, again attenuating the impact of frequency scales (for a discussion, see Menon, 1994; Menon et al., 1995).

Subsequent judgments. In addition to affecting respondents' behavioral reports, response alternatives may also influence subsequent judgments. Given respondents' assumption that the scale reflects the distribution of the behavior, checking a value on the scale amounts to determining one's location in the distribution, which influences subsequent comparative judgments. Accordingly, the patients in Schwarz and Scheuring's (1992) study of physical symptoms reported higher health satisfaction when the high-frequency scale suggested that their own symptom frequency is below average, relative to when the low-frequency scale suggested that it is above average. Note that this higher report of health satisfaction was obtained despite the fact that the former patients reported a higher symptom frequency in the first place, as seen above. Findings of this type arise because respondents extract comparison information from their own placement on the scale and use this information in making subsequent comparative judgments.

However, not all judgments are comparative in nature. When asked how satisfied we are with our health, we may compare our own symptom frequency to that of others. Yet, when asked how much our symptoms bother us, we may not engage in a social comparison but may instead draw on the absolute frequency of our symptoms. In this case, we may infer that our symptoms bother us more when a high frequency scale lead us to estimate a high symptom frequency. Accordingly, in another study, patients who reported their symptom frequency on one of the above scales reported that their symptoms bother them more when they received a high rather than a low frequency scale (Schwarz, 1999b). Thus, the same high-frequency scale elicited subsequent reports of higher health satisfaction (a comparative judgment) or of higher subjective suffering (a non-comparative judgment), depending on whether a comparative or a non-comparative judgment followed the symptom report.

Findings by Rothman, Haddock and Schwarz (in press) further illustrate the informational value of frequency scales for respondents. They asked undergraduates to report their number of sexual partners on a scale that presented either high or low numbers of partners. Subsequently, they assessed respondents' perception of HIV risk and their intention to use a condom. As expected, respondents drew on the comparison information provided by the scale in assessing their risk. Specifically, they inferred that they are at higher risk when their own (relatively high) placement on the low-frequency scale suggested that their own number of sexual partners is above average, relative to respondents whose own (relatively low) placement on the high-frequency scale suggested that their number of partners is below average.

Thus, they evaluated their risk by comparing their own number of partners to the "usual" number of partners suggested by the scale.

Yet, when asked about their intention to use a condom the next time they meet a new partner, respondents did not draw on their perception of their own sexual history, but on the information that the scale conveyed about the sexual history of others. Hence, they reported higher condom use intentions when the high-frequency scale suggested that others have numerous partners than when the low-frequency scale suggested that this is not the case.

In sum, frequency scales can influence subsequent judgments in different directions, depending on which source of information respondents draw on: their own relative placement on the scale, their own behavioral report, or the information conveyed about the likely behavior of others. These differential influences can result in apparently paradoxical answers, such as higher intention to use a condom despite lower perceived risk of HIV infection or higher health satisfaction despite higher symptom frequency and reports of being more bothered by one's physical symptoms.

Implications. Given the wide use of numeric response alternatives, it is worth highlighting the methodological implications.

First, numeric response alternatives influence respondents' interpretation of what the question refers to, as seen in the section on question comprehension. Hence, the same question stem, in combination with different frequency alternatives, may result in the assessment of differentially extreme behaviors.

Second, respondents' use of frequency scales as a frame of reference influences the obtained behavioral reports. Aside from calling the interpretation of the absolute values into question, this also implies that reports of the same behavior along different scales are not comparable, rendering comparisons between different studies difficult.

Third, because all respondents draw on the same frame of reference, frequency scales tend to homogenize the obtained reports. This reduces the observed variance as well as the likelihood that extreme groups are accurately identified. In evaluation studies, this may attenuate between-group differences in unwanted ways.

Fourth, the impact of response alternatives increases to the extent that respondents cannot recall relevant episodes from memory. This implies that reports of behaviors that are poorly represented in memory are more affected than reports of behaviors that are well represented (e.g., Menon et al., 1995). When differentially memorable behaviors are assessed, this may either exaggerate or attenuate any actual differences in the relative frequency of the behaviors, depending on the specific frequency range of the scale.

Fifth, for the same reason, respondents with poorer memory for the behavior under study are more likely to be influenced by response alternatives than are respondents with better memory. Such a differential impact of response alternatives on the reports provided by different groups of respondents can result in misleading conclusions about actual group differences (e.g., Ji, Schwarz, & Nisbett, 2000; Schwarz, 1999b).

Finally, the range of response alternatives may influence subsequent comparative and non-comparative judgments. Hence, respondents' may arrive at evaluative judgments that are highly context dependent and may not reflect the assessments they would be likely to make in daily life.

To avoid these systematic influences of response alternatives, it is advisable to ask frequency questions in an open response format, such as, "How many times a week do you. . . ? ___ times a week" Note that such an open format needs to specify the relevant units of measurement to

avoid answers such as "a few." Although the answers will be error prone because of the difficulty of accurate recall, they will at least not be systematically biased.

An Alternative? Vague Quantifiers

Given the difficulties associated with obtaining accurate quantitative frequency reports, along with the unintended side-effects of numeric frequency scales, researchers may be tempted to simplify respondents' task by using *vague quantifiers*, such as "sometimes," "frequently," and so on. If respondents cannot provide the desired details anyway, perhaps such global reports provide a viable alternative route? Unfortunately, this is not the case (for an extensive review, see Pepper, 1981).

Vague quantifiers do not reflect the absolute frequency of a behavior, but its frequency *relative* to the respondent's expectations. Hence, the same response (e.g., "sometimes") denotes different frequencies in different content domains and for different respondents. For example, "frequently" suffering from headaches reflects higher absolute frequencies than "frequently" suffering from heart attacks, undermining comparisons across different behaviors. Similarly, suffering from headaches "occasionally" denotes a higher frequency for respondents with a medical history of migraine than for those without, undermining comparisons across respondents.

These and related ambiguities (see Moxey & Sanford, 1992; Pepper, 1981) render vague quantifiers inadequate for the assessment of *objective* frequencies, despite the high popularity of their use. Instead, vague quantifiers provide an indirect assessment of the relationship between the frequency of a behavior and of respondents' expectations. If the latter information is of interest, it can be assessed in more direct ways.

Safeguarding Against Surprises

The undesirable influence of frequency scales is easily avoided by using an open-ended format, as discussed above. Respondents' reliance on subjective theories, on the other hand, presents a more complex problem. On the positive side, any inference requires some rudimentary "theory" about the content domain and, to the extent that respondents' theories are reasonably correct, theory-driven inferences may often be the best approximation we can get. On the negative side, there is no guarantee that respondents' subjective theories bear a close relationship to reality (Nisbett & Wilson, 1977; Ross, 1989). Most important in the context of evaluation research, the mere participation in any program is likely to evoke a subjective theory of change, or why else would one participate in it? This theory, in turn, may guide inferences that apparently confirm the expected changes. Worse, different groups, such as voluntary and involuntary participants, may rely on different theories, resulting in differential reports that suggest differential effectiveness. Again, cognitive pilot tests can alert researchers to participants' inference strategies and provide an opportunity to explore participants' subjective theories.

STEP 4: MAPPING THE ANSWER ONTO THE RESPONSE FORMAT

Once respondents have arrived at an answer in their own mind, they need to communicate it to the researcher. To do so, they may need to map their answer onto the response alternatives

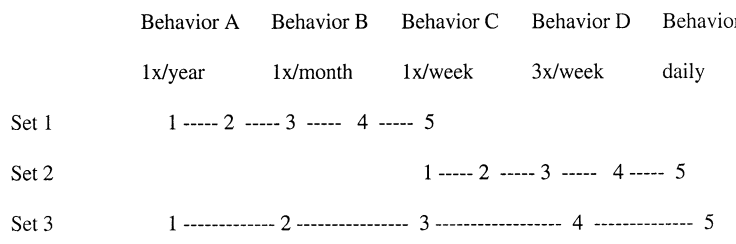

Figure 2. Anchoring rating scales.

provided by the researcher. For conceptual reasons, it is convenient to think of response formatting as a separate task. But in reality, response formatting is intimately intertwined with question comprehension and judgment formation, as we have already seen in the case of closed- versus open-ended questions or frequency scales. Here, we address respondents' use of rating scales and the emergence of response order effects in categorical questions.

Rating Scales

Rating scales are rarely used in the assessment of behavioral reports. When they are used, respondents are typically asked to rate the frequency of their behavior along a scale anchored by vague quantifiers, such as "rarely" or "very often." The problems that arise in this case resemble the problems discussed in the context of vague quantifiers.

First, when only one behavior is rated, respondents draw on their expectations to determine if its frequency qualifies as "very often," for example, as discussed above. Second, when several behaviors are rated along the same scale, the set of behaviors serves as a frame of reference that influences respondents' ratings. The underlying processes are conceptualized in Parducci's (1965) range-frequency theory, which we illustrate with an example.

Suppose that respondents are asked to rate the frequency of the behaviors shown in Fig. 2 along a 5-point scale, ranging from "sometimes" to "very often." To determine what qualifies as "very often," respondents attend to the range of behaviors they are asked to rate and anchor the rating scale with the least and most frequent behaviors on the list. If the list includes only behaviors A to C (Set 1), behavior C would receive a high rating, most likely a rating of 5. If the list includes only behaviors C to E (Set 2), behavior C would now appear as a low-frequency behavior in this context, most likely receiving a rating of 1 or 2. Finally, if the list included all behaviors A to E (Set 3), the scale would be "stretched" to accommodate the full range of frequencies. In this case, behavior C would most likely receive a rating of 3.

As this example illustrates, the frequency of a given behavior will be rated as less extreme if presented in the context of more frequent behaviors, than if presented in the context of less frequent ones. In Parducci's (1965) model, this impact of the range of behaviors is referred to as the *range effect.* In addition, if the number of behaviors to be rated is sufficiently large, respondents attempt to use all categories of the rating scale about equally often. Accordingly, the specific ratings given also depend on the number and distribution of the presented stimuli, an effect that is referred to as the *frequency effect.* Daamen and de Bie (1992) provide an introduction to the logic of these processes and report several studies that illustrate their impact on the obtained results.

Once again, the most important implication is that self-reports are context dependent and that reports obtained in different contexts cannot be directly compared. Suppose, for example, that an evaluator notices over the course of an intervention that the pretest questionnaire omitted some relevant, low-frequency behaviors. If these behaviors are included in the post-test questionnaire, thus changing the set of behaviors from something that resembled Set 2 to something that resembles Set 3 (Fig. 2), the ratings assigned to the original behaviors would shift, in this case suggesting that their frequency increased over the course of the intervention. Yet, all that happened might be a change in how respondents map their frequency estimates onto the rating scale provided by the researcher. It is, therefore, advisable not to change the set of behaviors whenever comparisons across time are of interest.

Response Order Effects

In many studies, respondents are presented with a list of behaviors they may have engaged in, or services they may have used, and are asked to check all that apply. The order in which the respective items are presented on the list may greatly influence the obtained reports. To date, the underlying processes have been more thoroughly explored for opinion questions than for behavioral questions (for a review and theoretical model see Sudman et al., 1996, chapter 6), and much remains to be learned about response order effects.

In general, a given item is more likely to be endorsed when it comes early rather than late on a list that is presented in a *visual* format (e.g., in a self-administered questionnaire or on a show card). For example, respondents of a German survey (reported in Schwarz, Hippler, & Noelle-Neumann, 1994) were asked, "Could you please tell me, with the help of this list, what you happened to do last Saturday?" The list presented 28 different activities, and the order in which these activities were listed was reversed for half of the respondents. Whereas 34% of the respondents reported that they worked on their job when this item was presented first, only 25% did so when this item was presented last. Conversely, 15% reported that they slept in when this item was presented first, whereas 10% reported doing so when this item was presented last. Note that these *primacy effects* were obtained in interviews conducted within 2 to 4 days after the Saturday in question, illustrating how fast the details of daily life are misremembered.

Several processes are likely to contribute to these effects (see Schwarz et al., 1994). First, respondents' effort to recall may decline over the course of the list because of fatigue. Accordingly, they may work harder at retrieving information pertaining to the first few items they consider. Second, the information brought to mind by earlier retrieval efforts may interfere with successful retrieval on subsequent items, thus limiting their endorsement. Third, respondents may feel that they have reported "enough" once they checked off a few items and may hence be less motivated to work on subsequent items. In some cases, such as the above example, the endorsement of an early item may also preclude the endorsement of a later one; having worked on one's job, for example, makes it less likely that one slept in.

Also important, the direction of response order effects *reverses* when the items are read to respondents (*auditory format*), rather than presented visually. In this case, a given item is more likely to be endorsed when it comes late rather than early on a list, resulting in a *recency effect.* To understand why, we need to consider the implications of visual and auditory presentation formats, as Krosnick and Alwin (1987) noted. When the list is presented in a visual format, respondents answer the items in the order in which they are presented,

presumably investing more effort in the early items. When the list is read to respondents, however, they have little opportunity to think about the early items because they need to pay attention to the additional material the interviewer is reading to them. Once the interviewer is done, however, respondents are likely to start with the items that are still "in their ears," that is, the last few read to them. This essentially reverses the process discussed above, giving an advantage to later items.

As a result of these diverging response order effects, responses to items presented in different orders, or under different presentation formats (visual vs. auditory), are again of limited comparability. Moreover, response order effects are particularly pronounced for older respondents because of their age-related limitations in the capacity of working memory (for a review, see Knäuper, 1999). In fact, older respondents may show response order effects under conditions where none are observed for younger respondents. Unfortunately, such age-sensitive context effects render comparisons across cohorts fraught with uncertainty (for reviews of other age-related differences in the response process, which exceed the scope of this article, see the contributions in Schwarz et al., 1999).

To safeguard against response order effects, researchers may choose one of two strategies. First, they may reverse the order in which the items are listed for half of the respondents. Although this ensures that the researcher becomes aware of possible response order effects, it remains unclear what to do with the results aside from the less than satisfying solution of averaging over both sets of answers. Alternatively, researchers can avoid the problem altogether by changing the format of the question. Instead of presenting a single list that asks respondents to "check all that apply," they can present each item as a separate yes/no question. The associated increase in interview time is relatively negligible, given that all items of the list need to be presented anyway, and the gain is worth the effort. Empirically, a format that requires a yes/no response to each item results in higher behavioral reports than an otherwise identical list format that requires respondents to "check all that apply" (e.g., Rasinski, Mingay, & Bradburn, 1994).

STEP 5: "EDITING" THE ANSWER

Social Desirability and Self-Presentation

Respondents' final task is to provide their answer to the interviewer, a stage at which they may decide to "edit" their answer for reasons of social desirability and self-presentation. Historically, a wide range of different context effects has been attributed to social desirability (for a review of the survey literature on this topic, see DeMaio, 1984), but recent research suggests that its influence is more limited. Nevertheless, respondents may deliberately provide inaccurate answers to threatening questions. A question is considered *threatening* when it pertains to a highly desirable or undesirable behavior. Respondents may find it embarrassing to admit that they did not engage in the desirable behavior or did engage in the undesirable one, resulting in over- and under-reporting, respectively. Also important, what is considered desirable or undesirable may often depend on the specific nature of the social situation: Whereas admitting that one has tried drugs may seem threatening to some teenagers when interviewed by an adult, admitting that one has never tried drugs may seem as threatening to some teens when interviewed by a peer. Moreover, respondents may be concerned that disclosing illegal behavior may have negative consequences.

Not surprisingly, socially desirable responding is more frequently observed in face-to-face interviews than in self-administered questionnaires, which provide a higher degree of confidentiality (e.g., Krysan, Schuman, Scott, & Beatty, 1994; Smith, 1979). All methods designed to reduce socially desirable responding focus on one of these two factors, question threat and confidentiality. To reduce question threat, researchers often embed threatening questions among less threatening ones (e.g., by presenting the target behavior on a list with more innocuous ones). Moreover, one may try to normalize the undesirable behavior along the lines of, "As you know, many people have been killing their spouses these days. Do you happen to have killed yours?" (for this and other examples, see Barton, 1958).

More promising than these attempts, however, are strategies designed to guarantee the privacy and confidentiality of respondents' answers. This is particularly important when personal interviews are conducted in a setting where other household members, or bystanders, can overhear the questions and answers. Of course, such settings are best avoided. If that is not feasible, the threatening question may be presented in writing and the respondent may return the answer in a sealed envelope, which has the additional advantage of maintaining the privacy of the response vis-à-vis the interviewer. A particularly elaborate version of this theme is known as the *randomized response technique* (e.g., Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; Horvitz, Shaw, & Simmons, 1967; Warner, 1965). In one variant of this technique, the respondent is given a card with two questions, one innocuous ("Where you born in April?") and one threatening ("Have you ever taken heroin?"). Both of these questions can be answered "yes" or "no" and which question the respondent is to answer is determined by a probability mechanism, such as drawing a colored bead from a box. The interviewer and researcher remain unaware which color the respondent has drawn, thus making it impossible to determine which question a particular respondent has answered. Knowledge of the proportion of different color beads in the box, however, allows researchers to estimate the proportion of respondents who have answered "yes" to the heroin question. As a result, this procedure permits estimates for the sample as a whole, but limits more detailed analyses because the responses cannot be linked to specific characteristics of the respondents.

Privacy concerns are somewhat less pronounced in telephone interviews, which have the advantage that bystanders cannot (usually) overhear the question. Hence, questions that allow for neutral answers (such as "yes/no," "strongly agree," or "twice") protect the respondent from disclosing sensitive information to bystanders. Unfortunately, researchers often fail to take advantage of this opportunity and present response alternatives that may be informative to bystanders even in the absence of knowledge of the question. Sudman and Bradburn (1983, chapter 3) provide detailed advice on the use of a wide range of question wording and confidentiality techniques, and we encourage readers to consult their suggestions.

Finally, a word of caution is appropriate. Based on what we said above, it may seem like a good idea to preempt any concerns respondents may have about the privacy of their answers by presenting detailed privacy and confidentiality assurances at the beginning of the interview and in an introductory letter that invites participation in a study. In fact, many researchers choose to take this route, which is often required by institutional review boards. Unfortunately, this strategy is likely to backfire. Prior to the actual interview, respondents cannot evaluate how threatening the questions might be. Given the maxims of conversation, they will infer from the researchers' assurances that they are likely to face embarrassing questions about sensitive topics, or why else would the researchers feel a need to provide these assurances? Once respondents see the actual questions, they may find them less

sensitive than the assurances suggested. But, unfortunately, many respondents may never see the questions, having decided not to participate when they pondered their likely nature (see Singer, Hippler, & Schwarz, 1992). To the extent possible, it is therefore preferable to introduce confidentiality assurances in low key terms at the initial contact stage and to provide specific confidentiality information at the relevant point in the interview, thus giving respondents an opportunity to make a truly informed decision.

CONCLUDING REMARKS

As this review indicates, self-reports of behavior can be profoundly influenced by the research instrument. At all steps of the response process—from question comprehension to recall, inference and estimation, response formatting and the eventually recorded answer—the information respondents provide depends in crucial ways on the specifics of the questionnaire. At first glance, these influences may seem more problematic for survey researchers than for evaluation researchers. Survey researchers want to arrive at an accurate estimate of the behavior in the population, whereas evaluation researchers primarily want to compare differences over time or between program participants and a control group. To the extent that features of question design affect participants in the same way at the pretest and post-test stage, or influence participants and control respondents in similar ways, little damage may be done. Although correct in principle, this hope is probably misleading in many cases.

As we noted throughout this review, many question design features *are* indeed likely to influence participants and control respondents in differential ways. At the question comprehension stage, participants may find behaviors noteworthy that "go without saying" for control respondents. At the recall stage, different behaviors may be memorable to participants and to control respondents. Both of these influences complicate comparisons between groups. Moreover, some behaviors may become more memorable over the course of an intervention that draws attention to them, complicating comparisons over time. At the inference and estimation stage, actual differences between groups, or over time, can be attenuated when all respondents rely on the same scale as a frame of reference in making an estimate. At the response formatting stage, more memorable behaviors are less likely to be subject to response order effects, again resulting in differential effects when memorability differs over time or between groups. Finally, when respondents report their answers to the interviewer, different groups may have differential reason to be concerned about confidentiality and social desirability. Throughout, these and other *differential* influences, as noted above, may distort comparisons over time or between groups, the very comparisons that are at the heart of evaluation research.

Unfortunately, there are no silver bullets of questionnaire design that assure accurate answers. Nor are there reliable cookbook recipes that work under all conditions. Instead, many design options come with their own specific tradeoffs, as we noted throughout this review. Hence, there is no alternative to thinking one's way through the specifics in each particular case. Despite these caveats, observation of a few simple points is likely to spare evaluators many headaches down the road:

First, answer every question yourself. If you find the task difficult, chances are that your respondents will find it next to impossible.

Second, your questionnaire is not a neutral instrument that merely collects information from respondents, but is also a source of information that respondents use to make sense of

the questions you ask (Schwarz, 1996). Hence, ask yourself what your respondents may infer from features of the questionnaire, including the response alternatives, the reference period, the content of related questions, the title of the questionnaire, and the sponsor of the study. Make sure that those features are consistent with the intended meaning of your questions.

Third, consult models of "good" questions. They can serve as useful starting points, but will usually require adjustment for the specific purpose at hand. We highly recommend Sudman and Bradburn's (1983) *Asking Questions* for this purpose.

Fourth, pilot test your questions with cognitive interviewing techniques that can alert you to the comprehension and recall problems that respondents encounter. Chapter 2 of Sudman et al.'s (1996) *Thinking About Answers* provides an introduction to these techniques, which can be employed with a small number of respondents from the target population. Adjust your questions and test them again.

Fifth, familiarize yourself with the basic psychology of asking and answering behavioral questions, to which this review provided an introduction. More comprehensive treatments can be found in *Thinking About Answers* (Sudman et al., 1996) and *The Psychology of Survey Response* (Tourangeau et al., 2000). An understanding of the basic principles is essential for appropriate questions and informed tradeoffs.

Sixth, encourage your respondents to invest the effort needed for providing accurate answers. Something as simple as acknowledging that the task is difficult, and instructing them that accuracy is important and that they should take all the time they need, can improve performance (for example instructions, see Cannell et al., 1981).

Seventh, where feasible, capitalize on the hierarchically nested structure of autobiographical memory by providing a meaningful context for respondents' memory search. Consider event history calendars as a possible format (see Belli, 1998).

Finally, ensure through interviewer training that your interviewers understand the intended meaning of your questions. Allow your interviewers to clarify questions when needed (Schober & Conrad, 1997) and make such clarifications part of the interviewer training.

We realize that some of these recommendations require a considerable commitment of time. But the time spent on good questionnaire design is a negligible cost in the overall budget of an evaluation study, and mistakes made at this stage cannot be corrected later on. The *GIGO* principle of "garbage in, garbage out" applies as much to evaluation research as to any other field. In the end, study results cannot be more meaningful than the raw data on which they are based. We, therefore, hope that the science underlying the collection of raw data will eventually figure as prominently in the training of evaluation researchers as the statistical techniques used to mine those data.

ACKNOWLEDGMENTS

This article was written while we were at the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. We gratefully acknowledge the Center's support and the helpful comments of several reviewers.

REFERENCES

Axinn, W., Pearce, L., & Ghimire, D. (1999). Innovations in life history calendar applications. *Social Science Research, 28,* 243–264.

Baddeley, A. (1990). *Human memory: Theory and practice.* Hillsdale, NJ: Erlbaum.

Barton, A. J. (1958). Asking the embarrassing question. *Public Opinion Quarterly, 22,* 271–278.

Belli, R. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory, 6,* 383–406.

Belli, R., Schwarz, N., Singer, E., & Talarico, J. (2000). Decomposition can harm the accuracy of retrospective behavioral reports. *Applied Cognitive Psychology, 14,* 295–308.

Belson, W. A. (1981). *The design and understanding of survey questions.* Aldershot: Gower.

Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research, 14,* 280–288.

Bradburn, N. M., Huttenlocher, J., & Hedges, L. (1994). Telescoping and temporal memory. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 203–216). New York: Springer Verlag.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236,* 157–161.

Brown, N. R. (in press). Encoding, representing, and estimating event frequencies: Multiple strategy perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition.* New York: Oxford University Press.

Cannell, C. F., Fisher, G., & Bakker, T. (1965). Reporting on hospitalization in the Health Interview Survey. *Vital and Health Statistics* (PHS Publication No. 1000, Series 2, No. 6). Washington, D.C.: US Government Printing Office.

Cannell, C. F., Miller, P. V., and Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological Methodology 1981* (pp. 389–437). San Francisco, CA: Jossey-Bass.

Caspi, A., Moffitt, T., Thornton, A., et al. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research, 6,* 101–114.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions* (pp. 15–48). New York: Russell Sage.

Collins, L. M., Graham, J. W., Hansen, W. B., & Johnson, C. A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. *Applied Psychological Measurement, 9,* 301–309.

Conrad, F. G., & Brown, N. R. (1996). Estimating frequency: A multiple strategy perspective. In D. Herrmann, M. Johnson, C. McEvoy, C. Hertzog, & P. Hertel (Eds.), *Basic and applied memory: Research on practical aspects of memory,* v Vol. 2 (pp. 167–178). Hillsdale, NJ: Erlbaum.

Conway, M. A. (1990). *Autobiographical memory: An introduction.* Buckingham, UK: Open University Press.

Daamen, D. D. L., & de Bie, S. E. (1992). Serial context effects in survey items. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 97–114). New York: Springer Verlag.

DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena,* Vol. 2 (pp. 257–281). New York: Russell Sage.

DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 177–196). San Francisco, CA: Jossey-Bass.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology, 66,* 633–645.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. In C. C. Clogg (Ed.) *Sociological Methodology,* Vol. 18 (pp. 37–68). Washington, D.C.: American Sociological Association.

Gaskell, G. D., O'Muircheartaigh, C. A., & Wright, D. B. (1994). Survey questions about the frequency of vaguely defined events: The effects of response alternatives. *Public Opinion Quarterly, 58,* 241–254.

Greenberg, B., Abul-Ela, A., Simmons, W., & Horvitz, D. (1969). The unrelated questions randomized response model theoretical framework. *Journal of the American Statistical Association, 64,* 421–426.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hippler, H. J., Schwarz, N., & Sudman, S. (Eds.). (1987). *Social information processing and survey methodology.* New York: Springer Verlag.

Horvitz, D. G., Shaw, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. In Hasho (Ed.), *Proceedings of the American Statistical Association* (pp. 65–72). Washington, D.C.: American Statistical Association.

Huttenlocher, J., Hedges, L. V., & Bradburn, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 196–213.

Jabine, T. B., Straf, M. L., Tanur, J. M., & Tourangeau, R. (Eds.). (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines.* Washington, D.C.: National Academy Press.

Jenkins, C. R., & Dillman, D. A. (1997). Towards a theory of self-administered questionnaire design. In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 165–196). New York: Wiley.

Ji, L., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin, 26,* 586–594.

Jobe, J., & Loftus, E. (Eds.). (1991). Cognitive aspects of survey methodology [Special issue]. *Applied Cognitive Psychology, 5.*

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior.* Morristown, NJ: General Learning Press.

Knäuper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly, 63,* 347–370.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51,* 201–219.

Krysan, M., Schuman, H., Scott, L. J., & Beatty, P. (1994). Response rates and response content in mail versus face-to-face surveys. *Public Opinion Quarterly, 58,* 381–399.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 259–292). San Francisco, CA: Jossey-Bass.

Levinson, S. C. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Linton, M. (1982). Transformations of memory in everyday life. In U. Neisser (Ed.), *Memory observed: Remembering in natural contexts* (pp. 77–91). San Francisco, CA: Freeman.

Loftus, E., & Fathi, D. C. (1985). Retrieving multiple autobiographical memories. *Social Cognition, 3,* 280–295.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? *Memory and Cognition, 11,* 114–120.

Lyketsos, C., Nestadt, G., Cwi, J. Heithoff, K., & Eaton, W. (1994). The life chart interview: A standardized method to describe the course of psychopathology. *International Journal of Methods in Psychiatric Research, 4,* 143–155.

Mathiowetz, N. A., & Duncan, G. J. (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. *Journal of Business and Economic Statistics, 6,* 221–229.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research, 20,* 431–440.

Menon, G. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 161–172). New York: Springer Verlag.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22,* 212–228.

Mingay, D. J., Shevell, S. K., Bradburn, N. M., & Ramirez, C. (1994). Self and proxy reports of everyday events. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 235–250). New York: Springer Verlag.

Moore, J. C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics, 4*(2), 155–172.

Moxey, L. M., & Sanford, A. J. (1992). Context effects and the communicative functions of quantifiers: Implications for their use in attitude research. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 279–296). New York: Springer Verlag.

Neisser, U. (1986). Nested structure in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–88). Cambridge, UK: Cambridge University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology, 29,* 1011–1020.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Park, J., Kosterman, R., Hawkins, D., Haggerty, K., Duncan, T., Duncan, S., et al. (2000). Effects of the "Preparing for the Drug Free Years" curriculum on growth in alcohol use and risk for alcohol use in early adolescence. *Prevention Science, 1,* 337–352.

Pepper, S. C. (1981). Problems in the quantification of frequency expressions. In D. W. Fiske (Ed.), *Problems with language imprecision,* New Directions for Methodology of Social and Behavioral Science, Vol. 9 (pp. 25–41). San Francisco, CA: Jossey-Bass.

Poulton, E. C. (1989). *Bias in quantifying judgments.* London, UK: Erlbaum.

Rasinski, K. A., Mingay, D., & Bradburn, N. M. (1994). Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly, 58,* 400–408.

Reiser, B. J., Black, J. B., & Abelson, R. P. (1985). Knowledge structure in the organization and retrieval of autobiographical memories. *Cognitive Psychology, 17,* 89–137.

Ross, M. (1989). The relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341–357.

Ross, M., & Conway, M. (1986). Remembering one's own past: The construction of personal histories. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 122–144). New York: Guilford.

Rothman, A. J., Haddock, G., & Schwarz, N. (in press). How many partners is too many? Shaping perceptions of personal vulnerability to HIV infection. *Journal of Applied Social Psychology*.

Schober, M. F. (1999). Making sense of questions: An interactional approach. In M. Sirken, D. Hermann, S. Schechter, N. Schwarz, J. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 77–94). New York: Wiley.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61,* 576–602.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. Zanna (Ed.), *Advances in experimental social psychology*, Vol. 26. San Diego, CA: Academic Press.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Hillsdale, NJ: Erlbaum.

Schwarz, N. (1999a). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

Schwarz, N. (1999b). Frequency reports of physical symptoms and health behaviors: How the

questionnaire determines the results. In D. C. Park, R. Morrell, & K. Shifren (Eds.), *Processing medical information in aging patients: Cognitive and human factors perspectives* (pp. 93–108). Mahway, NJ: Erlbaum.

Schwarz, N., & Hippler, H. J. (1991). Response alternatives: The impact of their choice and ordering. In P. Biemer, R. Groves, N. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 41–56). Chichester: Wiley.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly, 49,* 388–395.

Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1994). Retrospective reports: The impact of response alternatives. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 187–202). New York: Springer Verlag.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (Eds.). (1999). *Aging, cognition, and self-reports.* Washington, D.C.: Psychology Press.

Schwarz, N., & Scheuring, B. (1988). Judgments of relationship satisfaction: Inter- and intra-individual comparison strategies as a function of questionnaire structure. *European Journal of Social Psychology, 18,* 485–496.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Anwortvorgaben des Fragebogens lernen. (Frequency-reports of psychosomatic symptoms: What respondents learn from response alternatives.) *Zeitschrift für Klinische Psychologie, 22,* 197–208.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition, 6,* 107–117.

Schwarz, N., & Sudman, S. (Eds.). (1992). *Context effects in social and psychological research*. New York: Springer Verlag.

Schwarz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer Verlag.

Schwarz, N., & Sudman, S. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco, CA: Jossey-Bass.

Schwarz, N., & Wellens, T. (1997). Cognitive dynamics of proxy responding: The diverging perspectives of actors and observers. *Journal of Official Statistics, 13,* 159–179.

Sinclair, R. C., Mark, M. M., Moore, S. E., Lavis, C. A., & Soldat, A. S. (2000). An electoral butterfly effect. *Nature, 408,* 665–666.

Singer, E., Hippler, H. J., & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research, 4,* 256–268.

Sirken, M., Hermann, D., Schechter, S., Schwarz, N., Tanur, J., & Tourangeau, R. (Eds.). (1999). *Cognition and survey research*. New York: Wiley.

Skowronski, J. J., Betz, A. L., Thompson, C. P., Walker, W. R., & Shannon, L. (1994). The impact of differing memory domains on event-dating processes in self and proxy reports. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 217–234). New York: Springer Verlag.

Smith, T. W. (1979). Happiness. *Social Psychology Quarterly, 42,* 18–30.

Strube, G. (1987). Answering survey questions: The role of memory. In H. J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 86–101). New York: Springer Verlag.

Sudman, S., Bickart, B., Blair, J., & Menon, G. (1994). The effect of level of participation on reports of behavior and attitudes by proxy reporters. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 251–266). New York: Springer Verlag.

Sudman, S., & Bradburn, N. M. (1983). *Asking questions.* San Francisco, CA: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Sudman, S., & Schwarz, N. (1989). Contributions of cognitive psychology to advertising research. *Journal of Advertising Research, 29,* 43–53.

Tanur, J. M. (Ed.). (1992). *Questions about questions.* New York: Russell Sage.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions. The impact of data collection, mode, question format, and question context. *Public Opinion Quarterly, 60,* 275–304.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology, 18,* 225–252.

Wagenaar, W. A. (1988). People and places in my memory: A study on cue specificity and retrieval from autobiographical memory. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues,* Vol. 1 (pp. 228–232). Chichester: Wiley.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating error answer bias. *Journal of the American Statistical Association, 60,* 63–69.

Watson, D. (1982). The actor and the observer: How are their perceptions of causality divergent? *Psychological Bulletin, 92,* 682–700.

Whitten, W. B., & Leonard, J. M. (1981). Directed search through autobiographical memory. *Memory and Cognition, 9,* 566–579.

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long term memory. *Cognitive Science, 5,* 87–119.

Willis, G., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questions. *Applied Cognitive Psychology, 5,* 251–267.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of (emotion) frequency questions. *Journal of Personality and Social Psychology, 75,* 719–728.

Withey, S. B. (1954). Reliability of recall of income. *Public Opinion Quarterly, 18,* 31–34.