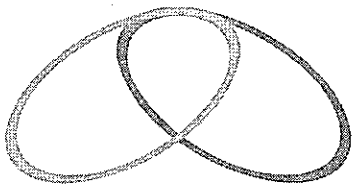
Social Psychology and Evaluation



EDITED BY

Melvin M. Mark
Stewart I. Donaldson
Bernadette Campbell

Chapter 9, p. 244-264

THE GUILFORD PRESS
New York London

To Kurt Lewin, Donald Campbell, Peter Rossi,
and their followers,
who toil at the intersection of social psychology and evaluation
in service of social betterment

A Division of Guilford Publications, Inc.
72 Spring Street, New York, NY 10012
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Social psychology and evaluation / edited by Melvin M. Mark, Stewart I. Donaldson, Bernadette Campbell.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-60918-212-0 (pbk.: alk. paper) — ISBN 978-1-60918-213-7 (hardcover: alk. paper)
1. Social psychology—Methodology. 2. Evaluation. I. Mark, Melvin M. II. Donaldson, Stewart I. (Stewart Ian) III. Campbell, Bernadette.
HM1033.S6423 2011
302.01—dc22

2011010212

CHAPTER 9

Asking Questions about Behavior

Self-Reports in Evaluation Research

Norbert Schwarz
Daphna Oyserman

Most interventions aim at changing people's behaviors. Accordingly, most evaluation studies include attempts to assess problem behaviors and to monitor their change over time. The questions used are usually straightforward, like these apparently simple questions about alcohol consumption: "Have you ever drunk beer, wine, wine coolers, whiskey, gin, or other liquor?" and, "How many times have you had beer, wine, or other= liquor in the past month?" These questions were adapted from Park and colleagues (2000), but similar questions can be found in many prevention studies and government health surveys in which participants are asked to self-report on the frequency of their behaviors in a specified period of time.

In posing such questions, researchers 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 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* accurately report on their behavior, 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, participants can rarely live up to these hopes.

This chapter illuminates why participants' can't deliver what researchers hope for and aims to help researchers to develop a more realistic approach. For this purpose, we introduce evaluation researchers to the cognitive and communicative processes underlying self-reports, drawing on extensive research at the interface of survey methodology and psychology (for comprehensive reviews see Sirken et al., 1999; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). Our discussion follows the key steps of the question-answering process and addresses (1) how respondents interpret the questions asked; (2) retrieve relevant information from memory; (3) draw inferences that allow them to move from the accessible information to a plausible answer; and (4) report the answer to the researcher. Throughout, our focus is on questions about behavior. For a discussion of questions about attitudes and evaluative judgments, including issues related to assessing program satisfaction, see Schwarz (2008) and Schwarz and Strack (1999). For a discussion of the implications of cultural differences on survey responses see Schwarz, Oyserman, and Peytcheva (2010).

Understanding the Question

As a first step, respondents need to understand the question to be able to provide a meaningful answer. Unfortunately, respondents' interpretation often fails to match what the researcher had in mind, even when the question is apparently simple and straightforward. For example, the expression "reading a magazine" has different meanings for different respondents, ranging from having seen it at the newsstand to having read it cover to cover (Belson, 1981). Hence, 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. To avoid such problems, textbooks urge researchers to avoid unfamiliar and ambiguous terms (see Bradburn, Sudman, & Wansink, 2004, for good advice). But ambiguities remain even when all terms are thoroughly familiar. When asked "What have you done today?" participants will understand the words, but they still need to determine what the researcher is interested in: Should they report, for example, that they took a shower or not? Merely understanding the words of a question, that is, its *literal meaning*, is not enough to answer it. Instead, an appropriate answer requires an

understanding of the *pragmatic meaning* of the question: What does the questioner want to know?

To infer what the questioner wants to know, participants draw on the tacit assumptions that underlie the conduct of conversations in everyday life (for reviews, see Clark & Schober, 1992; Schwarz, 1996). These assumptions can be summarized in the form of several conversational maxims (Grice, 1975). 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 expect respondents to interpret each question in isolation. This, however, is not what respondents do: They continue to interpret questions in context, and this gives rise to context effects in question interpretation. A *maxim of quantity* requests speakers to make their contribution as informative as is required, but not contribute more information than is required. This maxim invites respondents to provide information the questioner seems interested in, rather than other information that may come to mind or information that is inappropriate given the relationship between the two—or, what is colloquially termed "too much information." 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. Hence, 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 believe 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. While all participants are likely to be subtly influenced by immediate context, context sensitivity is likely to be higher in societies that are higher in collective cultural values and in contexts that cue collective mindsets (for details, see Oyserman, Coon, & Kemmelmeier, 2002; Oyserman, & Lee, 2007, 2008; Schwarz et al., 2010).

Open versus Closed Question Formats

With these 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. What kinds of 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 in this context and will report on them. In contrast, a control group of nonparticipants is unlikely to infer that the researcher is interested in brushing teeth, showering, or other grooming that they may interpret either as "things that go without saying" or as "too much information," and may therefore not report on the occurrence of 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, having been maintaining daily self-maintenance behaviors for a while may, find them less noteworthy than participants who just re-acquired these skills, raising additional comparison problems.

As an alternative approach, the evaluator may present a close-ended list of relevant behaviors. On the positive side, such a list would reduce the ambiguity of the open-ended question by indicating which behaviors are of interest, ensuring that control respondents report on behaviors that otherwise "go without saying." On the negative side, the list would also provide program participants with cues that may increase socially desirable responding. In addition, it may 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 close-ended question than when it needs to be volunteered in response to an open-ended question. At the same time, however, a close-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 close-ended question formats reliably result in different reports (see Schwarz & Hippler, 1991, for a review). On balance, a close-ended format is usually preferable if a reasonably complete list can be generated, although the tradeoffs need to be considered in each specific case.

Frequency Scales

Suppose the evaluator of an anger management program asks participants how frequently they felt "really irritated" or "got into a fight" recently. To answer this question, respondents have to determine what the researcher means by "really irritated" and "got into a fight." Does irritation refer to major or to minor annoyances? Are fights physical or verbal? To identify the intended meaning, participants may consult the *response alternatives* provided to them. 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 the idea that the researcher has relatively rare events

in mind. If so, respondents may conclude that the irritation 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 (see Schwarz, Strack, Müller, & Chassein, 1988, for experimental support). Thus, identically worded questions can acquire different meanings when accompanied by different frequency alternatives, leading respondents to report on substantively different behaviors.

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. Hence, comparisons across samples and sites cannot be made with confidence if the questions were not asked in identical ways—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 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. Supporting these predictions, Winkielman, Knäuper, and Schwarz (1998) observed that changes in the reference period shifted respondents' question interpretation and the resulting behavioral reports. It is therefore important to choose a reference period that is consistent with the intended meaning and to test respondents' interpretation at the questionnaire development stage, using the cognitive interviewing techniques we address in the next section.

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? Respondents may turn to adjacent questions for relevant cues. 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 like "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 meanings.

Safeguarding against Surprises

As the preceding examples illustrate, answering a question requires respondents to go beyond the literal meaning of the words to infer what the questioner wants to know. To do so, respondents turn to the context of the research conversation, much as they would be expected to in daily life (for a more detailed analysis see Schwarz, 1996). To safeguard against surprises at the question comprehension stage, we urge evaluators to look over their draft questionnaires and ask 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; Lessler and Forsyth (1996) offer an extensive checklist for this purpose. Once corrections have been made based on such a review, respondents' interpretation of questions can be explored in relatively inexpensive pilot tests using 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).

Recalling Relevant Information

Once respondents understand what they are to report on, they need to retrieve relevant information from 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. 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 (see Conrad & Brown, 1996, for a discussion). In fact, several factors render this strategy unsuitable for most of the behaviors of interest to evaluators.

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 one year after the event.

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 (e.g., Linton, 1982). This renders individual episodes of frequent behaviors indistinguishable and irretrievable.

Third, autobiographical knowledge is not organized by categories of behavior like "drinking alcohol" or the like. Instead, the structure of autobiographical memory can be thought of as a hierarchical network that includes *extended periods* (like "the years I lived in New York") at the highest level of the hierarchy. Nested within this high-order period are lower-level extended events pertaining to this time, like "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 unique. 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 (see Belli, 1998, for a comprehensive review). 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 will improve recall.

Researchers have developed a number of strategies that attempt to attenuate these problems, ranging from encouraging respondents to take their time to breaking a complex recall task into several simpler ones. Each of these strategies comes with specific advantages and disadvantages that require informed tradeoffs (see Schwarz & Oyserman, 2001, for a review and discussion). One noteworthy strategy with many helpful features and few drawbacks is the use of event history calendars.

Event History Calendars

Event history calendars are designed to take advantage of the nested structure of autobiographical memory (Belli, 1998). In a typical study, respondents are presented with a grid. The rows of the grid pertain to different aspects of their lives (e.g., where they lived, who they lived with, which jobs they had), and the columns represent time (e.g., years or months). Respondents reconstruct periods of their lives by completing this grid, going back and forth as successful recall of one aspect of their lives brings to mind other aspects that pertain to other parts of the grid. This structure supports the flexible use of multiple retrieval strategies within the hierarchically nested structure discussed above. Moreover, respondents are encouraged to take as much time as they need to complete the recall task and are told that accuracy is of great importance. Such encouragement reliably increases recall quality (see Cannell, Miller, & Oksenberg, 1981). Finally, event history calendars explicitly encourage the correction of earlier answers as newly recalled information qualifies earlier responses. Such correction opportunities are usually missed under regular interview formats, where respondents can 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 of the week 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.

Although the usefulness of event history calendars has been primarily demonstrated for the long-term recall of major events (like employment histories, criminal histories, or illness histories), the method can be adapted to shorter time periods and the assessment of more mundane behaviors (see Belli, 1998, for a review). While 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

In our experience, many recall questions would never be asked if researchers first tried to answer them themselves. Answering the questions you intend 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, inaccurate as it may be. It is therefore better to lower one's goals and design a more realistic, limited, and less demanding 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 (see Schwarz & Sudman, 1996). These techniques include, for example, a think aloud process in which participants are asked to say what they are thinking as they work on a question—what does each word mean, what comes to mind as they come up with an answer. More often than not, the experience will be sobering. 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.

Inference and Estimation

Given the difficulties associated with a recall-and-count strategy, it is not surprising that respondents usually resort to inference and estimation strategies to arrive at a plausible estimate (for reviews, see Conrad & Brown, 1996; Sudman et al., 1996, Ch. 9). We illustrate this with two particularly common strategies, which draw on subjective theories about the stability of one's behavior and on information provided by the questionnaire.

What My Behavior Must Have Been: Theory-Driven Inferences

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 to reflect the assumed change (see Ross, 1989, for a comprehensive review). 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.

In many domains, people 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. 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 pretraining 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).

As this discussion indicates, 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 ask-

ing the right question before the intervention, little can be done after the fact to make up for the oversight.

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. For example, a question about the frequency of physical symptoms may present one of these response scales:

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*

Consistent with the relevance maxim of conversation, participants assume that the researcher constructed a meaningful scale that is relevant to their task. Specifically, they assume that the values in the middle range of the scale correspond to the "average" or "usual" behavior and that the extremes of the scale correspond to the extremes of the distribution. As already seen in the section on question comprehension, the scale values are likely to influence respondents' interpretation of the question. In addition, they influence respondents' frequency estimates and related judgments.

Frequency Estimates

Respondents are likely to use the range and midpoint of the response alternatives as a frame of reference in estimating their own behavioral frequency. For example, they may infer that the midpoint represents the average or typical response and that the endpoints represent unusually high or low frequencies in the population. This means that responses are more likely to cluster at the mean if respondents assume that they are average. For example, when asked how much time they spent on homework using a close-ended scale, the average response of eighth graders in three inner-city schools did not differ from the midpoint of the scale, which represented 2 to 3 hours a week, but when the same children were asked the question without a scale, the average response was almost double, closer to 4 hours a week (Oyserman, 2009).

Moreover, in comparing responses to higher frequency and lower frequency scales, the above described estimation strategies result in higher frequency estimates along scales that present high- rather than low-frequency response alternatives. 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 above. 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. Overall, then, the two frequency scales resulted 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, sexual behaviors, and consumer behaviors (see Schwarz, 1999, for a review). As expected on theoretical grounds, the effect 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 (Menon, Raghubir, & Schwarz, 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 was below average, compared to when the low-frequency scale suggested their symptom frequency was 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 show that respondents extract comparison information from their own placement on the scale and use this information in making subsequent comparative judgments.

Not all judgments, however, are comparative in nature. When asked how satisfied we are with our health, we may compare our own symp-

tom 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 induced estimates of high symptom frequency. Accordingly, in another study, patients who reported their symptom frequency on one of the above scales reported that their symptoms bothered them more when they received a high- rather than a low-frequency scale (Schwarz, 1999). Thus, the same high-frequency scale elicited subsequent reports of higher health satisfaction (a comparative judgment) or of higher subjective suffering (a noncomparative judgment), depending on whether a comparative or a noncomparative judgment followed the symptom report.

Implications

Given the wide use of numeric response alternatives, it is worth highlighting the methodological implications of using this format. 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. This calls the interpretation of the absolute values into question and undermines the comparability of reports provided along different scales. 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. Fourth, the impact of response alternatives is more pronounced to the extent respondents cannot recall relevant episodes from memory. Hence, 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). This may 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. This can result in misleading conclusions about actual group differences (Schwarz, 2003). Sixth, the range of response alternatives may influence subsequent comparative and noncomparative 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. Finally there may be systematic cross-cultural differences in sensitivity to these effects and in the domains of sensitivity. This

may produce particularly misleading effects in studies involving participants of differing sociocultural backgrounds (Schwarz et al., 2010; Uskul, Oyserman, & Schwarz, 2010).

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." While the answers will be error prone, due to the difficulty of accurate recall, they will at least not be systematically biased by the frequency scale chosen by the evaluator.

An Alternative?: Vague Quantifiers

Given the reviewed difficulties, researchers may be tempted to simplify the respondents' task by using *vague quantifiers*, such as "sometimes" or "frequently." If respondents can't provide the desired details anyway, perhaps such global reports provide a viable alternative route? Unfortunately, this is not the case (see Pepper, 1981, for an extensive review).

Vague quantifiers do not reflect the absolute frequency of a behavior. Rather, they reflect 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, which undermines comparisons across different behaviors. Similarly, suffering from headaches "occasionally" denotes a higher frequency for respondents with a medical history of migraine than those without, which undermines comparisons across respondents.

These and related ambiguities (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 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 earlier. Respondents' reliance on subjective theories, on the other hand, presents a more complex problem. 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 (Ross, 1989). Worse, different groups—like 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.

Reporting the Answer: Confidentiality and Self-Presentation

Once respondents arrive at an answer in their own mind, they need to report it to the researcher. At this point, they may find it embarrassing to admit that they did not engage in a desirable behavior or did engage in an undesirable one, and they may "edit" their answers to ensure a more favorable self-presentation. Such deliberate misrepresentation of one's behavior is limited to highly threatening questions—that is, questions pertaining to highly desirable or undesirable behaviors (Bradburn et al., 2004; DeMaio, 1984). Importantly, what is considered desirable or undesirable may often depend on the specific nature of the social situation: While 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). Hence, it is important to guarantee the privacy and confidentiality of respondents' answers. Accordingly, settings in which other household members, or bystanders, can overhear the questions and answers 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. Bradburn et al. (2004) 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. Many researchers attempt to preempt respondents' possible confidentiality concerns 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, this 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 many respondents may never see the questions, having decided not to participate when they pondered their likely nature (see Singer, Hippler, & Schwarz, 1992). 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.

What to Do?

As this review indicates, self-reports of behavior and opinions can be profoundly influenced by the research instrument. Unfortunately, there are no silver bullets of questionnaire design that assure accurate answers. Instead, many design options come with their own specific tradeoffs, as we noted throughout this chapter. 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 nearly impossible.

Second, keep in mind that your questionnaire is not a neutral instrument that merely collects information from respondents. It 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 Bradburn and colleagues' (2004) *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. Pilot test your questions, adjust them—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 they should take all the time they need, can improve performance (see Cannell et al., 1981, for example, instructions).

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, 2002) and make such clarifications part of the interviewer training.

We realize that some of these recommendations require a considerable commitment of time in questionnaire development. Moreover, since answering questions well takes time, our recommendations also may require difficult choices as to what is really central to the evaluation (and so should be measured well) and what should not be attempted since it is unlikely that everything can be measured well. On the bright side, 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 in the training in the statistical techniques used to mine those data.

ACKNOWLEDGMENT

In preparing this chapter, we have drawn freely on Schwarz and Oyserman (2001); we thank the American Evaluation Association for permission to use this material.

REFERENCES

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.

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

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions* (2nd ed.). San Francisco: Jossey-Bass.

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, DC: U.S. Government Printing Office.

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

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* (Vol. 2, pp. 167–178). Hillsdale, NJ: Erlbaum.

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.

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.

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: Jossey-Bass.

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

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

Oyserman, D. (2009). *Self-reports of homework.* Unpublished data, Institute for Social Research, Ann Arbor, MI.

Oyserman, D., Coon, H., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128*, 3–73.

Oyserman, D., & Lee, S. S. W. (2007). Priming "culture": Culture as situated cognition. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 255–279). New York: Guilford Press.

Oyserman, D., & Lee, S. W. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin, 134*, 311–342.

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). San Francisco: Jossey-Bass.

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 Press.

Schober, M. F., & Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 67–94). New York: Wiley.

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

Schwarz, N. (1999). Frequency reports of physical symptoms and health behaviors: How the questionnaire determines the results. In D. C. Park, R. W. Morrell, & K. Shifren (Eds.), *Processing medical information in aging patients: Cognitive and human factors perspectives* (pp. 93–108). Mahwah, NJ: Erlbaum.

Schwarz, N. (2003). Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research, 29*, 588–594.

Schwarz, N. (2008). Attitude measurement. In W. Crano & R. Prislin (Eds.), *Attitudes and persuasion* (pp. 41–60). Philadelphia: Psychology Press.

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, UK: Wiley.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*, 127–160.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, et al. (Eds.), *survey methods in multinational, multiregional and multicultural contexts* (pp. 177–190). New York: Wiley.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten. (Frequency-reports of psychosomatic symptoms.) *Zeitschrift für Klinische Psychologie, 22,* 197–208.

Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61–84). New York: Russell Sage.

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. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research.* San Francisco: Jossey-Bass.

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.

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

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

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

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty or self-enhancement: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, et al. (Eds.), *Survey methods in multinational, multiregional and multicultural contexts* (pp. 191–202). New York: Wiley.

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.

EDITORS' CONCLUDING COMMENTS TO CHAPTER 9

Schwarz and Oyserman's review of the social and cognitive psychology research on self-reports can be a great help for evaluators who need to ask questions about behavior. You might want to ask yourself about the conditions under which self-reports of behavior are likely to be among the key measures in an evaluation (and conversely, when other kinds of measures will predominate). Keep in mind the ethical, financial, time, and other resource constraints that might preclude using other kinds of measures.

If you want to go more deeply into Schwarz and Oyserman's presentation, the following exercise may be useful. Consider an evaluation that you know about, or find a brief report of an evaluation (a search with the terms evaluation and journal will quickly locate a number of journals through which you can look). Did the evaluation include self-reports about behavior? If so, as best as you can tell, did the evaluators follow the suggestions laid out by Schwarz and Oyserman? Could they have done better? If the evaluation did not include self-reports about evaluation, why not, do you think? What measures of this type might they add? And how would they construct such a measure, according to the lessons of the chapter?

Like the other chapters in this section, this one has examined how social psychology can help address challenges that evaluators face in practice. There are many possibilities other than those presented in this section, and we hope that some readers will be motivated to think about the application of other areas of social psychology to evaluation.