

## **NOTICE CONCERNING COPYRIGHT RESTRICTIONS**

The copyright law of the United States [Title 17, United States Code] governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that use may be liable for copyright infringement.

The institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law. No further reproduction and distribution of this copy is permitted by transmission or any other means.

University of Chicago Interlibrary Loan  
OCLC: CGU / RLG:ILCG / DOCLINE: ILUJCL



ILLiad TN: 1241658

**Borrower:** EYM

**Lending String:** UPM,UPM,IAY,IUL,\*CGU

**Patron:** Oyserman, Daphna R  
**Reference #:** 7680551

**Journal Title:** International handbook of survey methodology /

**Volume: Issue:**  
**Month/Year:** 2008**Pages:** 18-34

EG

**Article Author:**

**Article Title:** Schwarz, N.; The psychology of asking questions

**Imprint:** New York ; London ; Lawrence Erlbaum Ass

**Notes:**

**ILL Number:** 76608507



**Call Number:** 4/18/2011 11:42:40 AM  
**Need By:**  
**Not Wanted After:** 05/15/2011  
**In Process:** 20110418  
**Received Via:**

**Notice:** This material may be protected by copyright law (Title 17 US Code)

2

**Call #:** HA31.2 .I565 2008 c.1

**Location:** Regenstein,  
**Bookstacks**

OCLC  
**Lanter**  
**Charge**  
**Maxcost:** \$26.00IFM  
Billing Notes:

**Shipping Address:**  
University of Michigan - Ann Arbor  
Hatcher Graduate Library - ILL  
913 S. University Ave.  
Ann Arbor, MI 48109-1205

**Fax:** 734-647-2050  
**Ariel:** ariel.lib.umich.edu  
**Email:** interlibrary.loan@umich.edu



**ODYSSEY REQUEST**

**SCAN for PLATES** (Pattern 2)

**EMAIL / ARIEL  
MAIL**

**SENT** \_\_\_\_\_

**Please report all Ariel transmission problems within 48 hours of receipt**

## Chapter 2

# The Psychology of Asking Questions

Norbert Schwarz  
*University of Michigan*

Bärbel Knäuper  
*McGill University*

Daphna Oyserman  
*University of Michigan*

Christine Stich  
*McGill University*

## 2.1 INTRODUCTION

Over the last two decades, psychologists and survey methodologists have made considerable progress in understanding the cognitive and communicative processes underlying survey responses, increasingly turning the “art of asking questions” (Payne, 1951) into an applied science that is grounded in basic psychological research. This chapter reviews key lessons learned from this work (for more extended reviews see Schwarz 1999a; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999; Sudman, Bradburn, & Schwarz 1996; Tourangeau, Rips, & Rasinski 2000). We focus on how features of the research instrument shape respondents’ answers and illustrate how the underlying processes can change as a function of respondents’ age and culture. We first address respondents’ tasks and subsequently discuss how respondents make sense of the questions asked. Next, we review how respondents answer behavioral questions and relate these questions to issues of autobiographical memory and estimation. Finally, we address attitude questions and review the conditions that give rise to context effects in attitude measurement.

## 2.2 RESPONDENTS’ TASKS

It is now widely recognized that answering a survey question involves several tasks. Respondents first need to understand the question to determine which information they are asked to provide. Next, they need to recall relevant information from memory. When the question is an opinion question, they will

rarely find a ready-for-use answer stored in memory. Instead, they need to form a judgment on the spot, based on whatever relevant information comes to mind at that time. When the question pertains to a behavior, respondents need to retrieve relevant episodes. Unless the behavior is rare and important, this is a difficult task and respondents typically have to rely on inference and estimation strategies to arrive at an answer. Once respondents have formed a judgment in their own minds, they can rarely report it in their own words. Instead, they need to format it to fit the response alternatives provided by the researcher. Finally, respondents may hesitate to communicate their private judgment, because of social desirability and self-presentation. If so, they may edit their judgment before conveying it to the researcher. Accordingly, understanding the question, recalling information, forming a judgment, formatting the judgment to fit the response alternatives, and editing the final answer are the major steps of the question answering process (see Strack & Martin, 1987; Tourangeau, 1984).

Unfortunately, respondents' performance at each of these steps is highly context dependent. From a psychological perspective, this context dependency is part and parcel of human cognition and communication, in daily life as in survey interviews. From a survey methods perspective, however, it presents a formidable problem: To the extent that the answers provided by the sample are shaped by the research instrument, they do not reflect the opinions or behaviors of the population to which the researcher wants to generalize. Complicating things further, a growing body of findings suggests that the underlying processes are age- and culture-sensitive, resulting in differential context effects that can thwart straightforward comparisons across cohorts and cultures.

### 2.3 UNDERSTANDING THE QUESTION

Survey textbooks typically advise researchers to avoid unfamiliar terms and complex syntax (for helpful guidelines see Bradburn, Sudman, & Wansink, 2004). This is good advice, but it misses a crucial point: Language comprehension is not about words per se, but about speaker meaning (Clark & Schober, 1992). Respondents certainly understand the words when asked, "What have you done today?" But to provide a meaningful answer they need to determine which behaviors the researcher might be interested in. For example, should they report that they took a shower, or not? To infer the intended meaning of the question, respondents rely on the tacit assumptions that govern the conduct of conversation in daily life. These assumptions were described by Paul Grice (1975), a philosopher of language, in the form of four maxims: A *maxim of relation* asks speakers to make their contribution relevant to the aims of the ongoing conversation. A *maxim of quantity* requests speakers to make their contribution as informative as is required, but not more informative than is required. A *maxim of manner* holds that a speaker's contribution should be clear rather than obscure, ambiguous or wordy, and a *maxim of quality* requires speakers not to say anything that's false. In short, speakers should try to be informative, truthful, relevant, and clear and listeners interpret the speakers' utterances "on the assumption that they are trying to live up to these ideals" (Clark & Clark, 1977, p. 122).

Respondents bring these tacit assumptions to the research situation and assume that the researcher "chose his wording so they can understand what he meant—and can do so quickly" (Clark & Schober, 1992, p. 27). To do so, they draw on the context of the ongoing conversation to determine the question's intended meaning, much as they would be expected to do in daily life. In fact, reliance on contextual information is more pronounced under the standardized conditions of survey interviews, where a well trained interviewer may merely reiterate the identical question, than under the less constrained conditions of daily life, which allow for mutual clarifications of the intended meaning. The contextual information provided by the researcher includes formal features of the questionnaire, in addition to the specific wording of the question and the content of preceding questions, as a few examples may illustrate (see Clark & Schober, 1992; Schwarz, 1996; Strack, 1994, for reviews).

### 2.3.1 Response Alternatives

Returning to the previously mentioned example, suppose respondents are asked in an *open response format*, "What have you done today?" To give a meaningful answer, they have to determine which activities may be of interest to the researcher. In an attempt to be informative, they are likely to omit activities that the researcher is obviously aware of (e.g., "I gave a survey interview") or may take for granted anyway (e.g., "I had breakfast"), thus observing the maxim of quantity. But most respondents would endorse these activities if they were included in a list presented as part of a *closed response format*. On the other hand, a closed response format would reduce the likelihood that respondents report any activities omitted from the list (see Schuman & Presser, 1981; Schwarz & Hippler, 1991, for reviews). This reflects that response alternatives convey what the researcher is interested in, thus limiting the range of "informative" answers. In addition, they may remind respondents of material that they may otherwise not consider.

Even something as innocuous as the *numeric values of rating scales* can elicit pronounced shifts in question interpretation. Schwarz, Knäuper, Hippler, Noelle-Neumann, and Clark (1991) asked respondents how successful they have been in life, using an 11-point rating scale with the endpoints labeled "not at all successful" and "extremely successful." To answer this question, respondents need to determine what is meant by "not at all successful"—the absence of noteworthy achievements or the presence of explicit failures? When the numeric values of the rating scale ranged from 0 to 10, respondents inferred that the question refers to different degrees of success, with "not at all successful" marking the absence of noteworthy achievements. But when the numeric values ranged from -5 to +5, with 0 as the middle alternative, they inferred that the researcher had a bipolar dimension in mind, with "not at all successful" marking the opposite of success, namely the presence of failure. Not surprisingly, this shift in the meaning of the verbal endpoint labels resulted in dramatic shifts in the obtained ratings. Whereas 34% of the respondents endorsed a value between 0 and 5 on the 0 to 10 scale, only 13% endorsed one of the formally equivalent values between -5 and 0 on the -5 to +5 scale, reflecting that the absence of great success is more common than the presence of failure. Hence, researchers are well advised to match the numeric values to the intended uni- or bipolarity of the scale.

The numeric values of behavioral *frequency scales* can serve a similar function. For example, Schwarz, Strack, Müller, and Chassein (1988) asked respondents to report how often they are angry along a scale that presented either high or low frequency values. As expected, respondents inferred that the question pertains to more intense anger experiences, which are relatively rare, when accompanied by low frequency values, but to mild anger experiences when accompanied by high frequency values. Throughout, respondents assume that the researcher constructs meaningful response alternatives that are relevant to the specific question asked, consistent with Grice's (1975) maxim of relation.

### 2.3.2 Question Wording

Similar issues apply to question wording. Minor changes in apparently formal features of the question can result in pronounced meaning shifts, as the case of *reference periods* may illustrate. Winkielman, Knäuper, and Schwarz (1998) asked respondents, in an open response format, either how frequently they had been angry last week or last year. Respondents inferred that the researcher is interested in less frequent and more severe episodes of anger when the question pertained to one year rather than to one week—after all, they could hardly be expected to remember minor anger episodes for a one-year period, whereas major anger may be too rare to make a one-week period plausible. Hence, they reported on rare and intense anger for the one year period, but more frequent and less intense anger for the one week period and their examples reflected this differential question interpretation. Accordingly, it is not surprising that reports across different reference periods do not add up—respondents may not even report on the same type of experience to begin with, thwarting comparisons across reference periods.

### 2.3.3 Question Context

Respondents' interpretation of a question's intended meaning is further affected by the context in which the question is presented. Hence, a question about drugs acquires a different meaning in the context of health versus a crime survey. Not surprisingly, the influence of *adjacent questions* is more pronounced for more ambiguously worded questions, which force respondents to rely on the context information to infer the intended meaning (e.g., Strack, Schwarz, & Wänke, 1991). Survey researchers have long been aware of this possibility (e.g., Payne, 1951). What is often overlooked, however, is that the *researcher's affiliation*, conveyed in the cover letter, may serve a similar function. For example, Norenzayan and Schwarz (1999) observed that respondents provided more personality focused explanations of a behavior when the questionnaire was printed on the letterhead of an "Institute for Personality Research" rather than an "Institute for Social Research." Such differences highlight the extent to which respondents as cooperative communicators attempt to make their answers relevant to the inferred epistemic interest of the researcher (see Schwarz, 1996).

### 2.3.4 Age-related Differences

Respondents' extensive use of contextual information requires that they hold the question in mind and relate it to other aspects of the questionnaire to determine its intended meaning. This entails considerable demands on respondents' cognitive resources. Given that these resources decline with increasing age (for a review see Park, 1999), we may expect that older respondents are less likely to use, or less successful in using, contextual information at the question comprehension stage. A limited body of findings supports this conjecture. For example, Schwarz, Park, Knäuper, Davidson, and Smith (1998) observed that older respondents (aged over 70) were less likely than younger respondents to draw on the numeric values of rating scales to interpret the meaning of endpoint labels. Similarly, Knäuper (1999a) observed in secondary analyses that question order effects decrease with age, as addressed in the section on attitude questions. Moreover, children and adolescents, whose cognitive capabilities are not yet fully developed, appear to show a similar deficit in incorporating relevant contextual information into survey responding (Borgers, de Leeuw, & Hox, 2000; Fuchs, 2005).

On theoretical grounds, age-related differences in the use of contextual information should be particularly likely in face-to-face and telephone interviews, where respondents can not look back to earlier questions. In contrast, they may be less pronounced in self-administered questionnaires, where respondents can deliberately return to previous questions when they encounter an ambiguous one (Schwarz & Hippler, 1995). If so, age-related differences in the response process may interact with the mode of data collection, further complicating comparisons across age groups.

### 2.3.5 Implications for Questionnaire Construction

As the preceding examples illustrate, question comprehension is not solely an issue of understanding the literal meaning of an utterance. Instead, it involves extensive inferences about the speaker's intentions to determine the pragmatic meaning of the question. To safeguard against unintended question interpretations and related complications, psychologists and survey methodologists have developed a number of procedures that can be employed in questionnaire pretesting (see Campanelli, chapter 10; Schwarz & Sudman, 1996). These procedures include the extensive use of probes and think-aloud protocols (summarily referred to as cognitive interviewing; e.g., DeMaio & Rothgeb, 1996), detailed coding of interview transcripts (e.g., Fowler & Cannell, 1996), and the use of expert systems that alert researchers to likely problems (e.g., Lessler & Forsyth, 1996). Without such development efforts, respondents' understanding of the questions asked may differ in important ways from what the researcher had in mind.

## 2.4 REPORTING ON ONE'S BEHAVIORS

Many survey questions pertain to respondents' behaviors, often asking them to report how frequently they engaged in a given behavior during a specified reference period. Ideally, respondents are supposed to determine the boundaries of the reference period and to recall all instances of the behavior within these boundaries to arrive at the relevant frequency. Unfortunately, respondents are usually unable to follow this recall-and-count strategy, unless the behavior is rare and important and the reference period short and recent (Menon, 1994). Instead, respondents will typically need to rely on estimation strategies to arrive at a plausible approximation. Next, we review key aspects of autobiographical memory and subsequently address respondents' estimation strategies.

### 2.4.1 Autobiographical Memory

Not surprisingly, people forget events in their lives as time goes by, even when the event is relatively important and distinct. For example, Cannell, Fisher, and Bakker (1965) observed that only 3% of their respondents failed to report an episode of hospitalization when interviewed within ten weeks of the event, yet a full 42% did so when interviewed one year after the event. Moreover, 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 (Linton, 1982; Strube, 1987). As a result, individual episodes of frequent behaviors become indistinguishable and irretrievable. Throughout, the available research suggests that the recall of individual behavioral episodes is largely limited to rare and unique behaviors of considerable importance, and poor even under these conditions.

Complicating things further, our autobiographical knowledge is not organized by categories of behavior (like drinking alcohol) that map easily onto survey questions. 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 earlier (e.g., "During that time, Lucy and I quarreled a lot"). Specific events, like a particular episode of 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 review). The search for any specific event in this network takes considerable time and the 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.



### 2.4.2 Facilitating Recall

Drawing on basic research into the structure of autobiographical memory, researchers have developed a number of strategies to facilitate autobiographical recall (for reviews see Schwarz & Oyserman, 2001; Sudman et al., 1996; Schwarz & Sudman, 1994; Tourangeau et al., 2000).

To some extent, researchers can improve the likelihood of accurate recall by restricting the recall task to a short and recent reference period. This strategy, however, may result in many zero answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with high behavioral frequencies. As a second strategy, researchers can provide appropriate recall cues. 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). Note, however, that recall cues share many of the characteristics of closed response formats and can constrain the inferred question meaning. It is therefore important to ensure that the recall cues are relatively exhaustive and compatible with the intended interpretation of the question.

Closely related to the provision of recall cues is the *decomposition* of a complex task into several more specific ones. Although this strategy results in reliable increases in reported frequency (e.g., Blair & Burton, 1987; Sudman & Schwarz, 1989), "more" is not always "better" and decomposition does not necessarily increase the accuracy of the obtained reports (e.g., Belli, Schwarz, Singer, & Talarico, 2000). As many studies documented, frequency estimates are regressive and people commonly overestimate low frequencies, but underestimate high frequencies (see Belli et al., 2000 for a review).

In addition, autobiographical recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds 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). Note, however, that it needs to be employed sparingly and may lose its credibility when used for too many questions within an interview.

Although the previously mentioned 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* (see Belli, 1998, for a review), which takes advantage of the hierarchically nested structure of autobiographical memory to facilitate recall. To help respondents recall their alcohol consumption during the last week, for example, they 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. 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 for recalling episodes of alcohol consumption.

### 2.4.3 Estimation Strategies

Given the reviewed memory difficulties, it is not surprising that respondents usually resort to a variety of inference strategies to arrive at a plausible estimate (for a review see Sudman et al., 1996, Chapter 9). Even when they can recall relevant episodic information, the recalled material may not cover the entire reference period or they 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 *decomposition* strategy (e.g., Blair & Burton, 1987). In other cases, respondents may draw on *subjective theories* that bear on the behavior in question (for a review see Ross, 1989). 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. Schwarz and Oyserman (2001) review these and related strategies. Here, we illustrate the role of estimation strategies by returning to respondents' use of information provided by formal characteristics of the questionnaire.

### 2.4.4 Response Alternatives

In many studies, respondents are asked to report their behavior by checking the appropriate response alternative on a *numeric frequency scale*. Consistent with Grice's (1975) maxim of relation, respondents assume that the researcher constructed a meaningful scale that is relevant to the task at hand. Specifically, they assume that values in the middle range of the scale reflect the average or "usual" behavior, whereas values at the extremes of the scale correspond to the extremes of the distribution. Given these assumptions, respondents can draw on the range of the response alternatives as a plausible frame of reference in estimating their own behavioral frequency. This results in higher frequency estimates when the scale presents high rather than low frequency values.

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 following two scales:

Low Frequency Scale	High Frequency Scale
<input type="checkbox"/> never	<input type="checkbox"/> twice a month or less
<input type="checkbox"/> about once a year	<input type="checkbox"/> once a week
<input type="checkbox"/> about twice a year	<input type="checkbox"/> twice a week
<input type="checkbox"/> twice a month	<input type="checkbox"/> daily
<input type="checkbox"/> more than twice a month	<input type="checkbox"/> several times a day

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. This influence of

frequency scales has been observed across a wide range of different behaviors, including health behaviors, television consumption (e.g., Schwarz, Hippler, Deutsch, & Strack, 1985), sexual behaviors (e.g., Tourangeau & Smith, 1996), and consumer behaviors (e.g., Menon, Rhaghubir, & Schwarz, 1995).

On theoretical grounds, we may expect that the impact of numeric frequency values is more pronounced, the more poorly the behavior is represented in memory, thus forcing respondents to rely on an estimation strategy. Empirically, this is the case. The influence of frequency scales is small when the behavior is rare and important, and hence well represented in memory. Moreover, when a respondent engages in the behavior with high regularity (e.g., every Sunday), its frequency can easily be derived from this rate information, largely eliminating the impact of frequency scales (Menon, 1994; Menon et al., 1995).

#### 2.4.5 Age- and Culture-related Differences in Estimation

Given age-related declines in memory, we may expect that the impact of response alternatives is more pronounced for older than for younger respondents. The available data support this prediction with some qualifications. For example, Knäuper, Schwarz, and Park (2004) observed that the frequency range of the response scale affected older respondents more than younger respondents when the question pertained to mundane behaviors, such as buying a birthday present. On the other hand, older respondents were less affected than younger respondents when the question pertained to the frequency of physical symptoms, which older people are more likely to monitor, resulting in better memory representations.

Similarly, Ji, Schwarz, and Nisbett (2000) observed pronounced cultural differences in respondents' need to estimate. In general, collectivist cultures put a higher premium on "fitting in" than individualist cultures (Oyserman, Coon, & Kimmelmeier, 2002). To "fit in," people need to monitor their own publicly observable behavior as well as the behavior of others to note undesirable deviations. Such monitoring is not required for private, unobservable behaviors. We may therefore expect that public behaviors are better represented in memory for people living in collectivistic rather than individualistic cultures, whereas private behaviors may be equally poorly represented in both cultures. To test these conjectures, Ji and colleagues (2000) asked students in China and the United States to report public and private behaviors along high or low frequency scales, or in an open response format. Replicating earlier findings, American students reported higher frequencies when presented with a high rather than low frequency scale, independent of whether the behavior was private or public. Chinese students' reports were similarly influenced by the frequency scale when the behavior was private, confirming that they relied on the same estimation strategy. In contrast, Chinese students' reports were unaffected by the response format when the behavior was public and hence needed to be monitored to ensure social fit.

As these examples illustrate, social groups differ in the extent to which they pay close attention to a given behavior. These differences in behavioral monitoring, in turn, influence to which extent respondents need to rely on estimation strategies in reporting on their behaviors, rendering them differentially susceptible to contextual influences. Importantly, such differences in respondents'

strategies can result in misleading substantive conclusions about behavioral differences across cultures and cohorts.

#### 2.4.6 Subsequent Judgments

In addition to affecting respondents' behavioral reports, frequency scales can also affect respondents' *subsequent judgments*. For example, respondents who check a frequency of twice a month on one of Schwarz and Scheuring's (1992) scales, shown earlier, may infer that their own symptom frequency is above average when presented with the low frequency scale, but below average when presented with the high frequency scale. Empirically, this is the case and the patients in this study reported higher health satisfaction after reporting their symptom frequencies on the high rather than low frequency scale – even though patients given a high frequency scale had reported a higher absolute symptom frequency to begin with. Again, such scale-induced comparison effects have been observed across a wide range of judgments (see Schwarz, 1999b for a review).

#### 2.4.7 Editing the Answer

After respondents arrived at an answer in their own mind, they need to communicate it to the researcher. At this stage, the communicated estimate may deviate from their private estimate due to considerations of social desirability and self-presentation as already mentioned (see DeMaio, 1984, for a review. Not surprisingly, editing on the basis of social desirability is particularly likely in response to threatening questions and is more pronounced in face-to-face interviews than in self-administered questionnaires, which provide a higher degree of confidentiality. All methods designed to reduce socially desirable responding address one of these two factors. Bradburn et al. (2004) review these methods and provide good advice on their use (see also Lensvelt-Mulders, Chapter 24).

#### 2.4.8 Implications for Questionnaire Construction

In sum, respondents will rarely be able to draw on extensive episodic memories when asked to report on the frequency of mundane behaviors. Instead, they need to rely on a variety of estimation strategies to arrive at a reasonable answer. Which strategy they use is often influenced by the research instrument, as the case of frequency scales illustrates. The most basic way to improve behavioral reports is to ensure that respondents have sufficient time to search memory and to encourage respondents to invest the necessary effort (Cannell et al., 1981). Moreover, it is usually advisable to ask frequency questions in an open response format, such as, "How many times a week do you ...? \_\_\_ times a week." Although the answers will not be accurate, the open response format will at least avoid the systematic biases associated with frequency scales.

Given these memory problems, researchers are often tempted to simplify the task by merely asking respondents if they engage in the behavior "never," "sometimes," or "frequently." Such *vague quantifiers*, however, are come with their own set of problems (see Pepper, 1981, for a review). For example,

"frequently" suffering from headaches reflects higher absolute frequencies than "frequently" suffering from heart attacks, and "sometimes" suffering from headaches denotes a higher frequency for respondents with a medical history of migraine than for respondents without that history. In general, the use of vague quantifiers reflects the objective frequency relative to respondents' subjective standard, rendering vague quantifiers inadequate for the assessment of objective frequencies, despite their popularity.

## 2.5 REPORTING ON ONE'S ATTITUDES

Public opinion researchers have long been aware that attitude measurement is highly context dependent. In this section, we address the two dominant sources of context effects in attitude measurement, namely the order in which questions and response alternatives are presented to respondents.

### 2.5.1 Question Order Effects

Dating back to the beginning of survey research, numerous studies demonstrated that preceding questions can influence the answers given to later questions (see Schuman & Presser, 1981; Schwarz & Sudman, 1992; Sudman et al., 1996; Tourangeau et al., 2000, for reviews). Moreover, when a self-administered questionnaire is used, respondents can go back and forth between questions, occasionally resulting in influences of later questions on responses to earlier ones (e.g., Schwarz & Hippler, 1995).

Question order effects arise for a number of different reasons. First, preceding questions can affect respondents' inferences about the intended meaning of subsequent questions, as discussed in the section on question comprehension (e.g., Strack, Schwarz, & Wänke, 1991). Second, they can influence respondents' use of rating scales, resulting in less extreme ratings when a given item is preceded by more extreme ones, which serve as scale anchors (e.g., Ostrom & Upshaw, 1968). Third, they can bring general norms to mind that are subsequently applied to other issues (e.g., Schuman & Ludwig, 1983). Finally, preceding questions can influence which information respondents use in forming a mental representation of the attitude object and the standard against which the object is evaluated.

The accumulating evidence suggests that a differential construal of attitude objects and standards is the most common source of question order effects. Hence, we focus on this aspect by following Schwarz and Bless' (1992a) inclusion/exclusion model, which predicts the direction and size of question order effects in attitude measurement, as well as their generalization across related issues.

### 2.5.2 Mental Construal

Attitude questions assess respondents' evaluations of an attitude object. From a psychological perspective, evaluations require two mental representations: A representation of the to-be-evaluated target and a representation of a standard,

against which the target is assessed. Both of these representations are formed on the basis of information that is accessible at the time of judgment. This includes information that may always come to mind when the respondent thinks about this topic (chronically accessible information), as well as information that may only come to mind because of contextual influences, for example information that was used to answer earlier questions (temporarily accessible information). Whereas temporarily accessible information is the basis of most context effects in attitude measurement, chronically accessible information lends some context-independent stability to respondents' judgments.

Independent of whether the information is chronically or temporarily accessible, people truncate the information search as soon as enough information has come to mind to form a judgment with sufficient subjective certainty. Hence, their judgment is rarely based on all information that may bear on the topic, but dominated by the information that comes to mind most easily at that point in time. How this information influences the judgment, depends on how it is used.

### 2.5.3 Assimilation Effects

Information that is *included* in the temporary representation formed of the target results in *assimilation effects*. That is, including information with positive implications results in a more positive judgment, whereas including information with negative implications results in a more negative judgment. For example, Schwarz, Strack, and Mai (1991) asked respondents to report their marital satisfaction and their general life-satisfaction in different question orders. When the general life-satisfaction question was asked first, it correlated with marital satisfaction  $r = .32$ . Reversing the question order, however, increased this correlation to  $r = .67$ . This reflects that the marital satisfaction question brought marriage related information to mind that respondents included in the representation formed of their lives in general. Accordingly, happily married respondents reported higher general life-satisfaction in the marriage-life than in the life-marriage order, whereas unhappily married respondents reported lower life-satisfaction under this condition.

As this pattern indicates, the specific effect of thinking about one's marriage depends on whether it is a happy or unhappy one. Accordingly, no overall mean difference was observed for the sample as a whole, despite pronounced differences in correlation. As a general principle, question order effects are not a function of the preceding question per se, but of the information that the question brings to mind. Hence, pronounced question order effects may occur in the absence of overall mean differences, rendering measures of association more sensitive than examinations of means.

Theoretically, the size of assimilation effects increases with the amount and extremity of the temporarily accessible information, and decreases with the amount and extremity of chronically accessible information, that is included in the representation of the target (e.g., Bless, Schwarz, & Wänke, 2003). To continue with the previously mentioned example, some respondents were asked to report on their job satisfaction, leisure satisfaction, and marital satisfaction prior to reporting on their general life-satisfaction, thus bringing a more varied range of information about their lives to mind. As expected, this decreased the correlation

of marital satisfaction and general life-satisfaction from  $r = .67$  to  $r = .43$ . By the same token, we expect that respondents who are experts on a given issue show less pronounced assimilation effects than novices, because experts can draw on a larger set of chronically accessible information, which in turn reduces the impact of adding a given piece of temporarily accessible information. Note, however, that expert status needs to be defined with regard to the specific issue at hand. Global variables, such as years of schooling, are unlikely to moderate the size of assimilation effects, unless they are confounded with the amount of knowledge regarding the issue under consideration. Accordingly, formal education has been found to show inconsistent relationships with the emergence and size of question order effects (Schuman & Presser, 1981).

#### 2.5.4 Contrast Effects

What has long rendered the prediction of question order effects challenging, is that the same piece of information that elicits an assimilation effect may also result in a *contrast effect*. This is the case when the information is excluded from, rather than included in, the cognitive representation formed of the target (Schwarz & Bless, 1992a). As a first possibility, suppose that a given piece of information with positive (negative) implications is excluded from the representation of the target. If so, the representation contains less positive (negative) information, resulting in a less positive (negative) judgment. For example, the Schwarz et al. (1991) life-satisfaction study included a condition in which the marital satisfaction and life-satisfaction questions were introduced with a joint lead-in that read, "We now have two questions about your life. The first pertains to your marriage and the second to your life in general." This lead-in was designed to evoke the conversational maxim of quantity (Grice, 1975), which enjoins speakers to avoid redundancy when answering related questions. Accordingly, respondents who had just reported on their marriage should now disregard this aspect of their lives when answering the general life-satisfaction question. Confirming this prediction, happily married respondents now reported lower general life-satisfaction, whereas unhappily married respondents reported higher life-satisfaction, indicating that they excluded the positive (negative) marital information from the representation formed of their lives in general. These diverging effects reduced the correlation to  $r = .18$ , from  $r = .67$  when the same questions were asked in the same order without a joint lead-in. Finally, a control condition in which the general life-satisfaction question was reworded to, "Aside from your marriage, which you already told us about, how satisfied are you with your life in general?" resulted in a highly similar correlation of  $r = .20$ . Such *subtraction based* contrast effects are limited to the specific target (here, one's life in general), reflecting that merely subtracting a piece of information (here, one's marriage) does only affect this specific representation. The size of subtraction based contrast effects increases with the amount and extremity of the temporarily accessible information that is excluded from the representation of the target, and decreases with the amount and extremity of the information that remains in the representation of the target.

As a second possibility, respondents may not only exclude accessible information from the representation formed of the target, but may also use this information in constructing a standard of comparison. If the implications of the

temporarily accessible information are more extreme than the implications of the chronically accessible information used in constructing a standard, this process results in a more extreme standard, eliciting contrast effects for that reason. The size of these comparison based contrast effects increases with the extremity and amount of temporarily accessible information used in constructing the standard or scale anchor, and decreases with the amount and extremity of chronically accessible information used in making this construction. In contrast to subtraction based comparison effects, which are limited to a specific target, comparison based contrast effects generalize to all targets to which the standard is applicable.

As an example, consider the impact of political scandals on assessments of the trustworthiness of politicians. Not surprisingly, thinking about a politician who was involved in a scandal, say Richard Nixon, decreases trust in politicians in general. This assimilation effect reflects that the exemplar is included in the representation formed of the target politicians in general. If the trustworthiness question pertains to a specific politician, however, say Bill Clinton, the primed exemplar cannot be included in the representation formed of the target—after all, Bill Clinton is not Richard Nixon. In this case, Richard Nixon may serve as a standard of comparison, relative to which Bill Clinton seems very trustworthy. Experiments with German exemplars confirmed these predictions (Schwarz & Bless, 1992b; Bless, Igou, Schwarz, & Wänke, 2000): Thinking about a politician who was involved in a scandal decreased the trustworthiness of politicians in general, but increased the trustworthiness of all specific exemplars assessed. In general, the same information is likely to result in assimilation effects in the evaluation of superordinate target categories (which allow for the inclusion of all information pertaining to subordinate categories), but in contrast effects in the evaluation of lateral target categories (which are mutually exclusive).

### 2.5.5 Determinants of Inclusion/Exclusion

Given the crucial role of inclusion/exclusion operations in the construction of mental representations, it is important to understand their determinants. When thinking about a topic, people generally assume that whatever comes to mind bears on what they are thinking about—or why else would it come to mind now? Hence, the default information is to include information that comes to mind in the representation of the target. This renders assimilation effects more likely than contrast effects. In fact, assimilation effects (sometimes referred to as carry-over effects) dominate the survey literature and many models intended to account for question order effects don't even offer a mechanism for the conceptualization of contrast effects (e.g., Zaller, 1992), which severely limits their usefulness as general theoretical frameworks. Whereas inclusion is the more common default, the exclusion of information needs to be triggered by salient features of the question answering process. The most relevant variables can be conceptualized as bearing on three implicit decisions that respondents have to make with regard to the information that comes to mind.

Some information that comes to mind may simply be irrelevant, pertaining to issues that are unrelated to the question asked. Other information may potentially be relevant to the task at hand and respondents have to decide what to do with it. The first decision bears on why this information comes to



mind. Information that seems to come to mind for the wrong reason, for example because respondents are aware of the potential influence of a preceding question, is likely to be excluded. The second decision bears on whether the information that comes to mind bears on the target of judgment or not. The content of the context question (e.g., Schwarz & Bless, 1992a), the superordinate or lateral nature of the target category (e.g., Schwarz & Bless, 1992b), the extremity of the information (e.g., Herr, 1986), or its representativeness for the target category (e.g., Strack, Schwarz, & Gschneidinger, 1985) are relevant at this stage. Finally, conversational norms of nonredundancy may elicit the exclusion of previously provided information, as seen earlier (Schwarz et al., 1991).

Whenever any of these decisions results in the exclusion of information from the representation formed of the target, it will elicit a contrast effect. Whether this contrast effect is limited to the target, or generalizes across related targets, depends on whether the excluded information is merely subtracted from the representation of the target or used in constructing a standard against which the target is evaluated. Whenever the information that comes to mind is included in the representation formed of the target, on the other hand, it results in an assimilation effect. Hence, the inclusion/exclusion model provides a coherent conceptualization of the emergence, direction, size, and generalization of context effects in attitude measurement (see Schwarz & Bless, 1992a; Sudman et al., 1996, Chapter 5, for more detail).

### 2.5.6 Age- and Culture-related Differences

To guard against question order effects, survey researchers often separate related questions with buffer items. These buffer items presumably render the previously used information less accessible, thus attenuating the influence of earlier questions (for a review see Wänke & Schwarz, 1997). The same logic suggests that preceding questions should be less likely to influence the judgments of older respondents, due to age-related declines in memory. Empirically this is the case, as Knäuper (1999a) observed in secondary analyses of survey data.

Much as age-related differences in memory performance can elicit age-sensitive context effects, culture-related differences in conversational practice can elicit culture-sensitive context effects. For example, Asian cultures value more indirect forms of communication, which require a higher amount of reading between the lines, based on high sensitivity to subtle conversational cues. Accordingly, Asians are more likely to notice the potential redundancy of related questions, as Haberstroh, Oyserman, Schwarz, Kühnen and Ji (2002) observed in a conceptual replication of the previously mentioned marital satisfaction study (Schwarz et al., 1991) with Chinese respondents. Throughout, such age- and culture-sensitive context effects can invite misleading conclusions about age- and culture-related differences in respondents' attitudes.

### 2.5.7 Response Order Effects

Another major source of context effects in attitude measurement is the order in which response alternatives are presented. Response order effects are most

reliably obtained when a question presents several plausible response options (see Sudman et al., 1996, chapter 6, for a detailed discussion). Suppose, for example, that respondents are asked in a self-administered questionnaire whether divorce should be easier to obtain or more difficult to obtain. When they first think about the easier option, they may quickly come up with a good reason for making divorce easier and may endorse this answer. But had they first thought about the more difficult option, they might as well have come up with a good reason for making divorce more difficult and might have endorsed that answer. In short, the order in which response alternatives are presented can influence the mental representation that respondents form of the issue (see Sudman et al., 1996, for a more detailed discussion).

Which response alternative respondents are more likely to elaborate on first, depends on the presentation order and mode (Krosnick & Alwin, 1987). In a visual format, like a self-administered questionnaire, respondents think about the response alternatives in the order in which they are presented. In this case, a given alternative is more likely to be endorsed when presented first rather than last, resulting in a *primacy effect*. In an auditory format, like a telephone interview, respondents cannot think about the details until the interviewer has read the whole question. In this case, they are likely to begin with the last alternative read to them, which is still in their ear. Under this format, a given alternative is more likely to be endorsed when presented last rather than first, resulting in a *recency effect*.

### 2.5.8 Age-related Differences

On theoretical grounds, we may expect that age-related limitations of working memory capacity further enhance respondents' tendency to elaborate mostly on a single response alternative. Empirically this is the case and an extensive meta-analysis documented that response order effects are more pronounced for older and less educated respondents (Knäuper, 1999b). This age-sensitivity of response order effects can again invite misleading conclusions about cohort differences in the reported attitude, suggesting, for example, that older respondents are more liberal than younger respondents under one order condition, but more conservative under the other (Knäuper, 1999a).

The observation that response order effects increase with age, whereas question order effects decrease with age, also highlights that age-sensitive context effects do indeed reflect age-related differences in cognitive capacity, which can plausibly account for both observations. In contrast, attempts to trace these differences to age-related differences in attitude strength (e.g., Sears, 1986) would suggest that question order and response order effects show parallel age patterns, which is not the case.

### 2.5.9 Implications for Questionnaire Construction

Human judgment is always context dependent, in daily life as in survey interviews. Although attention to the theoretical principles summarized earlier can help researchers to attenuate context effects in attitude measurement, the best safeguard against misleading conclusions is the experimental variation of question and response order within a survey.

## 2.6 CONCLUDING REMARKS

Survey researchers have long been aware that collecting data by asking questions is an exercise that may yield many surprises. Since the 1980s, psychologists and survey methodologists have made considerable progress in understanding the cognitive and communicative processes underlying question answering, rendering some of these surprises less surprising than they have been in the past. Yet, this does not imply that we can always predict how a given question would behave when colleagues ask us for advice: In many cases, the given question is too mushy an operationalization of theoretical variables to allow for predictions (although we typically feel we know what would happen if the question were tinkered with, in one way or another, to bring it in line with theoretical models). Nevertheless, the accumulating insights (reviewed in Sudman et al., 1996; Tourangeau et al., 2000) alert us to likely problems and help us in identifying questions and question sequences that need systematic experimental testing before they are employed in a large-scale study.

## GLOSSARY OF KEY CONCEPTS

**Assimilation effect.** A catch-all term for any influence that makes the answers to two questions more similar than they otherwise would be; it does not entail specific assumptions about the underlying process.

**Backfire effect.** See contrast effect.

**Carry-over effect.** See assimilation effect.

**Context effect.** A catch-all term for any influence of the context in which a question is asked; it does not entail specific assumptions about the direction of the effect or the underlying process.

**Contrast effect.** A catch-all term for any influence that makes the answers to two questions more different than they otherwise would be; it does not entail specific assumptions about the underlying process.

**Pragmatic meaning.** Refers to the intended (rather than literal or semantic) meaning of an utterance and requires inferences about the speaker's knowledge and intentions.

**Primacy effect.** A given response alternative is more likely to be chosen when presented at the beginning rather than at the end of a list of response alternatives.

**Question order effect.** The order in which questions are asked influences the obtained answers; different processes can give rise to this influence.

**Recency effect.** A given response alternative is more likely to be chosen when presented at the end rather than at the beginning of a list of response alternatives.

**Response order effect.** The order in which response alternatives are presented influences which alternative is endorsed; see primacy effect and recency effect.

**Semantic meaning.** Refers to the literal meaning of words. Understanding the semantic meaning is insufficient for answering a question, which requires an understanding of the question's pragmatic meaning.