RESEARCH REPORT SERIES
(Survey Methodology #2006-13)

Survey Questionnaire Construction

Elizabeth Martin

Director's Office
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: December 21, 2006

Survey Questionnaire Construction

Elizabeth Martin
U. S. Census Bureau, Washington D.C.

Glossary

closed question A survey question that offers response categories.
context effects The effects that prior questions have on subsequent responses.
open question A survey question that does not offer response categories.
recency effect Overreporting events in the most recent portion of a reference period, or a tendency to select the last-presented response alternative in a list.
reference period The period of time for which a respondent is asked to report.
response effects The effects of variations in question wording, order, instructions, format, etc. on responses.
retention interval The time between an event to be remembered and a recall attempt.
screening questions Questions designed to identify specific conditions or events.
split-sample An experimental method in which a sample is divided into random subsamples and a different version of a questionnaire is assigned to each.
standardized questionnaire The wording and order of questions and response choices are scripted in advance and administered as worded by interviewers.

Questionnaires are used in sample surveys or censuses to elicit reports of facts, attitudes, and other subjective states. Questionnaires may be administered by interviewers in person or by telephone, or they may be self-administered on paper or another medium, such as audio-cassette or the internet. Respondents may be asked to report about themselves, others in their household, or other entities, such as businesses. This article focuses on construction of standardized survey questionnaires.

The utility of asking the same questions across a broad group of people in order to obtain comparable information from them has been appreciated at least since 1086, when William the Conqueror surveyed the wealth and landholdings of England using a standard set of inquiries and compiled the results in the "Domesday Book." Sophistication about survey techniques has increased vastly since then, but fundamental insights about questionnaires advanced less during the millennium than might have been hoped. For the most part, questionnaire construction has remained more an art than a science. In recent decades there have been infusions of theory from relevant disciplines (such as cognitive psychology and linguistic pragmatics), testing and evaluation techniques have grown more comprehensive and informative, and knowledge about questionnaire design effects and their causes has cumulated. These developments are beginning to transform survey questionnaire construction from an art to a science.

Theoretical Perspectives on Asking and Answering Questions

Three theoretical perspectives point toward different issues that must be considered in constructing a questionnaire.

The Model of the Standardized Survey Interview

From this perspective, the questionnaire consists of standardized questions that operationalize the measurement constructs. The goal is to present a uniform stimulus to respondents so that their responses are comparable. Research showing that small changes in question wording or order can substantially affect responses has reinforced the assumption that questions must be asked exactly as worded, and in the same order, to produce comparable data.

Question Answering as a Sequence of Cognitive Tasks

A second theoretical perspective was stimulated by efforts to apply cognitive psychology to understand and perhaps solve recall and reporting errors in surveys of health and crime. A respondent must perform a series of cognitive tasks in order to answer a survey question. He or she must comprehend and interpret the question, retrieve relevant information from memory, integrate the information, and respond in the terms of the question. At each stage, errors may be introduced. Dividing the response process into components has provided a framework for exploring response effects, and has led to new strategies for questioning. However, there has been little research demonstrating that respondents actually engage in the hypothesized sequence of cognitive operations when they answer questions, and the problems of retrieval that stimulated the application of cognitive psychology to survey methodology remain nearly as difficult as ever.

The Interview as Conversation

Respondents do not necessarily respond to the literal meaning of a question, but rather to what they infer to be its intended meaning. A survey questionnaire serves as a script performed as part of an interaction between respondent and interviewer. The interaction affects how the script is enacted and interpreted. Thus, the construction of meaning is a social process, and is not carried by question wording alone. Participants in a conversation assume it has a purpose, and rely upon implicit rules in a cooperative effort to understand and achieve it. They take common knowledge for granted and assume that each participant will make his contribution relevant and as informative as required, but no more informative than necessary. (These conversational maxims were developed by Paul Grice, a philosopher). The resulting implications for the interview process are:

1. Asking a question communicates that a respondent should be able to answer it.
2. Respondents interpret questions to make them relevant to the perceived intent.
3. Respondents interpret questions in ways that are relevant to their own situations.
4. Respondents answer the question they think an interviewer intended to ask.
5. Respondents do not report what they believe an interviewer already knows.
6. Respondents avoid providing redundant information.
7. If response categories are provided, at least one is true.

These implications help us understand a number of well-established questionnaire phenomena. Consistent with item 1, many people will answer survey questions about unfamiliar objects using the question wording and context to construct a plausible meaning. As implied by items 2 and 3, interpretations of questions vary greatly among respondents. Consistent with item 4, postinterview studies show that respondents do not believe the interviewer "really" wants to know everything that might be reported, even when a question asks for complete reports. Consistent with items 5 and 6, respondents reinterpret questions to avoid redundancy. As implied by item 7, respondents are unlikely to volunteer a response that is not offered in a closed question.

The conversational perspective has been the source of an important critique of standardization, which is seen as interfering with the conversational

resources that participants would ordinarily employ to reach a common understanding, and it has led some researchers to advocate flexible rather than standardized questioning. A conversational perspective naturally leads to a consideration of the influences that one question may have on interpretations of subsequent ones, and also the influence of the interview context–what respondents are told and what they infer about the purposes for asking the questions–on their interpretations and responses.

Constructing Questionnaires

Constructing a questionnaire involves many decisions about the wording and ordering of questions, selection and wording of response categories, formatting and mode of administration of the questionnaire, and introducing and explaining the survey. Although designing a questionnaire remains an art, there is increasing knowledge available to inform these decisions.

Question Wording

Although respondents often seem to pay scant attention to survey questions or instructions, they are often exquisitely sensitive to subtle changes in words and syntax. Question wording effects speak to the power and complexity of language processing, even when respondents are only half paying attention.

A famous experiment illustrates the powerful effect that changing just one word can have in rare cases. In a national sample, respondents were randomly assigned to be asked one of two questions:

1. "Do you think the United States should allow public speeches against democracy?"
2. "Do you think the United States should forbid public speeches against democracy?"

Support for free speech is greater–by more than 20 percentage points–if respondents answer question 2 rather than question 1. That is, more people answer "no" to question 2 than answer "yes" to question 1; "not allowing" speeches is not the same as "forbidding" them, even though it might seem to be the same. The effect was first found by Rugg in 1941 and later replicated by Schuman and Presser in the United States and by Schwarz in Germany in the decades since, so it replicates in two languages and has endured over 50 years–even as support for freedom of speech has increased, according to both versions.

Terminology

"Avoid ambiguity" is a truism of questionnaire design. However, language is inherently ambiguous, and seemingly simple words may have multiple meanings.. Research by Belson and others demonstrates that ordinary words and phrases, such as "you," "children," and "work,"are interpreted very differently by different respondents.

Complexity and Ambiguity

Both cognitive and linguistic factors may impede respondents' ability to understand a question *at all*, as well as give rise to variable or erroneous interpretations. Questionnaire designers often intend a survey question to be interpreted literally. For example:

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

Such questions challenge respondents who must parse the question, interpret its key referents (i.e., "doctor or assistant," "medical doctor of any kind"), infer the events to be included (visits to discuss respondent's health in person or by telephone during the past 12 months) and excluded (visits while in a hospital), and keep in mind all these elements while formulating an answer. Apart from a formidable task of recall, parsing such a complex question may overwhelm available mental resources so that a respondent does not understand the question fully or at all. Processing demands are increased by embedded clauses or sentences (e.g., "while you were a patient in a hospital") and by syntactic ambiguity. An example of syntactic ambiguity appears in an instruction on a U. S. census questionnaire to include "People living here

most of the time while working, even if they have another place to live." The scope of the quantifier "most" is ambiguous and consistent with two possible interpretations, (i) "...[most of the time][while working]..." and (ii) "... [most of the [time while working]]...."

Ambiguity also can arise from contradictory grammatical and semantic elements. For example, it is unclear whether the following question asks respondents to report just one race: "I am going to read you a list of race categories. Please choose one or more categories that best indicate your race." "One or more" is contradicted by the singular reference to "race" and by "best indicate," which is interpretable as a request to select one.

Cognitive overload due to complexity or ambiguity may result in portions of a question being lost, leading to partial or variable interpretations and misinterpretations. Although the negative effects of excessive burden on working memory are generally acknowledged, the practical limits for survey questions have not been determined, nor is there much research on the linguistic determinants of survey question comprehension.

Presupposition

A presupposition is true regardless of whether the statement itself is true or false–that is, it is constant under negation. (For example, the sentences "I am proud of my career as a survey methodologist" and "I am not proud of my career as a survey methodologist" both presuppose I have a career as a survey methodologist.) A question generally shares the presuppositions of its assertions. "What are your usual hours of work?" presupposes that a respondent works, and that his hours of work are regular. Answering a question implies accepting its presuppositions, and a respondent may be led to provide an answer even if its presuppositions are false. Consider an experiment by Loftus in which subjects who viewed accident films were asked "Did you see *a* broken headlight?" or "Did you see *the* broken headlight?" Use of the definite article triggers the presupposition that there was a broken headlight, and people asked the latter question were more likely to say "yes," irrespective of whether the film showed a broken headlight.

As described by Levinson, linguists have isolated a number of words and sentence constructions that trigger presuppositions, such as change of state verbs (e.g., "Have you *stopped* attending church?"), and factive verbs (e.g., "regret," "realize," and "know"). (For example, "If you knew that the AMA is opposed to Measure H, would you change your opinion from *for* Measure H to *against* it?" presupposes the AMA is opposed to Measure H.) Forced choice questions, such as "Are you a Republican or a Democrat?" presuppose that one of the alternatives is true.

Fortunately for questionnaire designers, presuppositions may be cancelled. "What are your usual hours of work?" might be reworded to ask, "What are your usual hours of work, or do you not have usual hours?" Filter questions [e.g., "Do you work?" and (if yes) "Do you work regular hours?"] can be used to test and thereby avoid unwarranted presuppositions.

Question Context and Order

Question order changes the context in which a particular question is asked. Prior questions can influence answers to subsequent questions through several mechanisms. First, the semantic content of a question can influence interpretations of subsequent questions, especially when the subsequent questions are ambiguous. For example, an obscure "monetary control bill" was more likely to be supported when a question about it appeared after questions on inflation, which presumably led respondents to infer the bill was an anti-inflation measure.

Second, the thoughts or feelings brought to mind while answering a question may influence answers to subsequent ones. This is especially likely when an answer to a question creates expectations for how a subsequent one should be answered. A famous experiment manipulated the order of a pair of questions:

"Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?"

"Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?"

Respondents were much more likely to think

Communist reporters should be allowed in the United States if they answered that question second. Respondents apparently answered whichever question was asked first in terms of pro-American or anti-Communist sentiments. The second question activated a norm of reciprocity. Since many respondents felt constrained to treat reporters from both countries equally, they gave an answer to the second question that was consistent with the first.

Third, following conversational maxims, respondents may interpret questions so they are not redundant with prior questions. When a specific question precedes a general question, respondents "subtract" their answer to the specific question from their answer to the general one, under certain circumstances. Respondents asked questions about marital satisfaction and general life satisfaction reinterpret the general question to exclude the specific one: "Aside from your marriage, which you already told us about, how satisfied are you with other aspects of your life?"

This type of context effect, called a part–whole effect by Schuman and Presser, can occur for factual as well as attitudinal questions. For example, race and Hispanic origin items on the U. S. census form are perceived as redundant by many respondents, although they are officially defined as different. When race (the more general item) appears first, many Hispanic respondents fail to find a race category with which they identify, so they check "other" and write in "Hispanic." When Hispanic origin is placed first so that such respondents first have a chance to report their Hispanic identity, they are less likely to report their Hispanic origin in the race item. Thus, when the specific item comes first, many respondents reinterpret race to exclude the category Hispanic. In this case, manipulating the context leads to reporting that is more consistent with measurement objectives.

One might wonder why a prior question about marital satisfaction would lead respondents to exclude, rather than include, their feelings about their marriages in their answers to a general life satisfaction question. Accounts of when information primed by a prior question will be subtracted rather than assimilated into later answers or interpretations have been offered by Schwarz and colleagues and by Tourangeau *et al.*

The argument is that when people are asked to form a judgment they must retrieve some cognitive representation of the target stimulus, and also must determine a standard of comparison to evaluate it. Some of what they call to mind is influenced by preceding questions and answers, and this temporarily accessible information may lead to context effects. It may be added to (or subtracted from) the representation of the target stimulus. The questionnaire format and the content of prior questions may provide cues or instructions that favor inclusion or exclusion. For example, Schwarz and colleagues induced either an assimilation or a contrast effect in German respondents' evaluations of the Christian Democratic party by manipulating a prior knowledge question about a highly respected member (X) of the party. By asking "Do you happen to know which party X has been a member of for more than twenty years?" respondents were led to add their feelings about X to their evaluation of the party in a subsequent question, resulting in an assimilation effect. Asking "Do you happen to know which office X holds, setting him aside from party politics?" led them to exclude X from their evaluation of the party, resulting in a contrast effect.

Alternatively, the information brought to mind may influence the standard of comparison used to judge the target stimulus and result in more general context effects on a set of items, not just the target. For example, including Mother Teresa in a list of public figures whose moral qualities were to be evaluated probably would lower the ratings for everyone else on the list. Respondents anchor a scale to accommodate the range of stimuli presented to them, and an extreme (and relevant) example in effect shifts the meaning of the scale. This argues for explicitly anchoring the scale to incorporate the full range of values, to reduce such contextual influences.

Response Categories and Scales

The choice and design of response categories are among the most critical decisions about a questionnaire. As noted, a question that offers a choice among alternatives presupposes that one of

them is true. This means that respondents are unlikely to volunteer a response option that is not offered, even if it might seem an obvious choice.

Open versus Closed Questions

An experiment by Schuman and Presser compared open and closed versions of the question, "What do you think is the most important problem facing this country at present?" The closed alternatives were developed using responses to the open-ended version from an earlier survey. Just as the survey went in the field, a prolonged cold spell raised public fears of energy shortage. The open version registered the event: "food and energy shortages" responses were given as the most important problem by one in five respondents. The closed question did not register the energy crisis because the category was not offered in the closed question, and only one respondent volunteered it.

This example illustrates an advantage of open questions, their ability to capture answers unanticipated by questionnaire designers. They can provide detailed responses in respondents' own words, which may be a rich source of data. They avoid tipping off respondents as to what response is normative, so they may obtain more complete reports of socially undesirable behaviors. On the other hand, responses to open questions are often too vague or general to meet question objectives. Closed questions are easier to code and analyze and compare across surveys.

Types of Closed-Response Formats

The previous example illustrates that response alternatives must be meaningful and capture the intended range of responses. When respondents are asked to select only one response, response alternatives must also be mutually exclusive.

The following are common response formats:

Agree–disagree: Many survey questions do not specify response alternatives but invite a "yes" or "no" response. Often, respondents are offered an assertion to which they are asked to respond: for example, "Do you agree or disagree?–Money is the most important thing in life." Possibly because they state only one side of an issue, such items encourage acquiescence, or a tendency to agree regardless of content, especially among less educated respondents.

Forced choice: In order to avoid the effects of acquiescence, some methodologists advocate explicitly mentioning the alternative responses. In a stronger form, this involves also providing substantive counterarguments for an opposing view:

"If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law?"

"If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law because it would be too difficult to enforce?"

Formal balance, as in the first question, does not appear to affect response distributions, but providing counterarguments does consistently move responses in the direction of the counterarguments, according to Schuman and Presser's experiments. Devising response options with counterarguments may not be feasible if there are many plausible reasons for opposition, since the counterargument can usually only capture one.

Ordered response categories or scales: Respondents may be asked to report in terms of absolute frequencies (e.g., "Up to ½ hour, ½ to 1 hour, 1 to 1 ½ hours, 1 ½ to 2 hours, 2 to 2 ½ hours, More than 2 ½ hours"), relative frequencies (e.g., "All of the time, most of the time, a good bit of the time, some of the time, a little bit of the time, none of the time"), evaluative ratings (e.g., "Excellent, pretty good, only fair, or poor"), and numerical scales (e.g., "1 to 10" and "-5 to +5").

Response scales provide a frame of reference that may be used by respondents to infer a normative response. For example, Schwarz and colleagues compared the absolute frequencies scale presented in the previous paragraph with another that ranged from "Up to 2 ½ hour" to "More than 4 ½ hours" in a question asking how many hours a day the respondent watched television. The higher scale led to much higher frequency reports, presumably because many respondents were influenced by what they perceived to be the normative or average (middle) response in the scale. If there is a strong normative expectation, an open-ended question may avoid this source of bias. Frequently, ordered categories are intended to measure where a respondent belongs on an

underlying dimension (scale points may be further assumed to be equidistant). Careful grouping and labeling of categories is required to ensure they discriminate. Statistical tools are available to evaluate how well response categories perform. For example, an analysis by Reeve and Mâsse (see Presser *et al.*) applied item response theory to show that "a good bit of the time" in the relative frequencies scale presented previously was not discriminating or informative in a mental health scale.

Rating scales are more reliable when all points are labeled and when a branching structure is used, with an initial question (e.g., "Do you agree or disagree...") followed up by a question inviting finer distinctions ("Do you strongly agree/disagree, or somewhat agree/disagree?"), according to research by Krosnick and colleagues and others. The recommended number of categories in a scale is 7, plus or minus 2. Numbers assigned to scale points may influence responses, apart from the verbal labels. Response order may influence responses, although the basis for primacy effects (i.e., selecting the first category) or recency effects (i.e., selecting the last category) is not fully understood. Primacy effects are more likely with response options presented visually (in a self-administered questionnaire or by use of a show card) and recency effects with aural presentation (as in telephone surveys).

Offering an Explicit "Don't Know" Response Option

Should "don't know" be offered as an explicit response option? On the one hand, this has been advocated as a way of filtering out respondents who do not have an opinion and whose responses might therefore be meaningless. On the other hand, it increases the number of respondents who say "don't know," resulting in loss of data. Schuman and Presser find that the relative proportions choosing the substantive categories are unaffected by the presence of a "don't know" category, and research by Krosnick and others suggests that offering "don't know" does not improve data quality or reliability. Apparently, many respondents who take the easy out by saying "don't know" when given the opportunity are capable of providing meaningful and valid responses. Thus, "don't know" responses are best discouraged.

Communicating Response Categories and the Response Task

Visual aids, such as show cards, are useful for communicating response categories to respondents in personal interviews. In self-administered questionnaires, the categories are printed on the questionnaire. In either mode, the respondent does not have to remember the categories while formulating a response, but can refer to a printed list. Telephone interviews, on the other hand, place more serious constraints on the number of response categories; an overload on working memory probably contributes to the recency effects that can result from auditory presentation of response options. Redesigning questions to branch, so that each part involves a smaller number of options, reduces the difficulty. Different formats for presenting response alternatives in different modes may cause mode biases; on the other hand, the identical question may result in different response biases (e.g., recency or primacy effects) in different modes. Research is needed on this issue, especially as it affects mixed mode surveys.

The same general point applies to communicating the response task. For example, in developmental work conducted for implementation of a new census race question that allowed reports of more than one race, it proved difficult to get respondents to notice the "one or more" option. One design solution was to introduce redundancy, so respondents had more than one chance to absorb it.

Addressing Problems of Recall and Retrieval

Psychological theory and evidence support several core principles about memory that are relevant to survey questionnaire construction:

1. Autobiographical memory is reconstructive and associative.

2. Autobiographical memory is organized hierarchically. (Studies of free recall suggest the organization is chronological, with memories for specific events embedded in higher order event sequences or periods of life.)

3. Events that were never encoded (i.e., noticed, comprehended, and stored in memory) cannot be recalled.

4. Cues that reinstate the context in which an event was encoded aid memory retrieval.

5. Retrieval is effortful and takes time.

6. Forgetting increases with the passage of time due to decay of memory traces and to interference from new, similar events.

7. The characteristics of events influence their memorability: salient, consequential events are more likely to be recalled than inconsequential or trivial ones.

8. Over time, memories become less idiosyncratic and detailed, and more schematic and less distinguishable from memories for other similar events.

9. The date an event occurred is usually one of its least accurately recalled features.

Principle 6 is consistent with evidence of an increase in failure to report events, such as hospitalizations or consumer purchases, as the time between the event and the interview–the retention interval–increases. Hospitalizations of short duration are more likely to be forgotten than those of long duration, illustrating principle 7. A second cause of error is telescoping. A respondent who recalls that an event occurred may not recall when. On balance, events tend to be recalled as happening more recently than they actually did–that is, there is forward telescoping, or events are brought forward in time. Forward telescoping is more common for serious or consequential events (e.g., major purchases and crimes that were reported to police). Backward telescoping, or recalling events as having happened longer ago than they did, also occurs. The aggregate effect of telescoping and forgetting is a pronounced recency bias, or piling up of reported events in the most recent portion of a reference period. Figure 1 illustrates the effect for two surveys.

The rate for the month prior to the interview is taken as a base and the rates for other months are calculated relative to it. Line 3 shows that monthly victimization rates decline monotonically each month of a 6-month reference period. Lines 1 and 2 show the same for household repairs over a 3-month reference period; note the steeper decline for minor repairs.

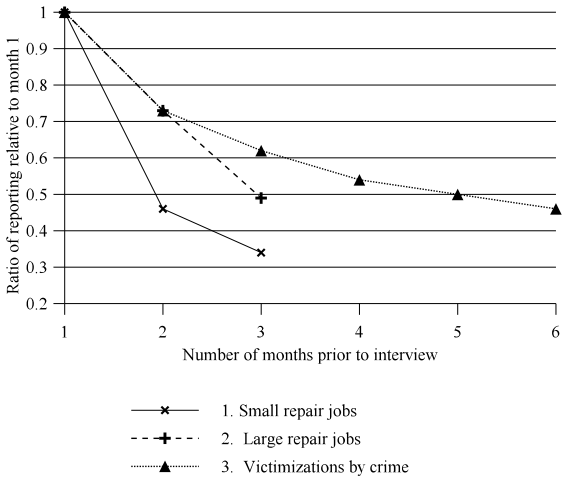

Figure 1 Recency bias for two surveys. *Sources:* Neter, J. and Waxberg, J. (1964) "A Study of Response Errors in Expenditures Data from Household Interviews." *J. Am. Stat. Assoc.* 59:18-55; Biderman, A. D. and Lynch, J. P. (1981) "Recency Bias in Data on Self-Reported Victimization" *Proc. Social Stat. Section (Am. Stat. Assoc.)*: 31-40.

Recent theories explain telescoping in terms of an increase in uncertainty about the timing of older events. Uncertainty only partially explains telescoping, however, since it predicts more telescoping of minor events than of major ones, but in fact the opposite occurs.

Because of the serious distortions introduced by failure to recall and by telescoping, survey methodologists are generally wary of "Have you ever...?"-type questions that ask respondents to recall experiences over a lifetime. Instead, they have developed various questioning strategies to try to improve respondents' recall.

Strategies to Improve Temporal Accuracy

In order to improve recall accuracy, questions are usually framed to ask respondents to recall events that occurred during a reference period of definite duration. Another procedure is to bound an interview with a prior interview, in order to prevent respondents from telescoping in events that happened before the reference period. Results of the bounding interview are not included in survey estimates. Another method attempts to make the boundary of the reference period more vivid by associating it with personal or historical landmark events. This can reduce telescoping, especially if the landmark is relevant to the types of events a

respondent is asked to recall. A more elaborate procedure, the event history calendar, attempts to structure flexible questions in a way that reflects the organization of memory and has proved promising in research by Belli and associates.

For many survey questions, respondents may rely on a combination of memory and judgment to come up with answers. When the number of events exceeds 10, very few respondents actually attempt to recall and enumerate each one. Instead, they employ other strategies, such as recalling a few events and extrapolating a rate over the reference period, retrieving information about a benchmark or standard rate and adjusting upward or downward, or guessing. By shortening the reference period, giving respondents more time, or decomposing a question into more specific questions, questionnaire designers can encourage respondents to enumerate episodes if that is the goal.

Aided and Unaided Recall

In general, unaided (or free) recall produces less complete reporting than aided recall. It may also produce fewer erroneous reports. Cues and reminders serve to define the scope of eligible events and stimulate recall of relevant instances. A cuing approach was employed to improve victimization reporting in a 1980s redesign of the U. S. crime victimization survey. Redesigned screening questions were structured around multiple frames of reference (acts, locales, activities, weapons, and things stolen), and included numerous cues to stimulate recall, including recall for underreported, sensitive, and nonstereotypical crimes. The result was much higher rates of reporting.

Although cuing improves recall, it can also introduce error, because it leads to an increase in reporting of ineligible incidents as well as eligible ones. In addition, the specific cues can influence the kinds of events that are reported. The crime survey redesign again is illustrative. Several crime screener formats were tested experimentally. The cues in different screeners emphasized different domains of experience, with one including more reminders of street crimes and another placing more emphasis on activities around the home. Although the screeners produced the same overall rates of victimization, there were large differences in the characteristics of crime incidents reported. More street crimes and many more incidents involving strangers as offenders were elicited by the first screener.

Dramatic cuing effects such as this may result from the effects of two kinds of retrieval interference. Part-set cuing occurs when specific cues interfere with recall of noncued items in the same category. For example, giving “knife” as a weapons cue would make respondents less likely to think of “poison” or “bomb” and (by inference) less likely to recall incidents in which these noncued items were used as weapons. The effect would be doubly biasing if (as is true in experimental studies of learning) retrieval in surveys is enhanced for cued items and depressed for noncued items.

A second type of interference is a retrieval block that occurs when cues remind respondents of details of events already mentioned rather than triggering recall of new events. Recalling one incident may block retrieval of others, because a respondent in effect keeps recalling the same incident. Retrieval blocks imply underreporting of multiple incidents. Early cues influence which event is recalled first, and once an event is recalled, it inhibits recall for additional events. Therefore, screen questions or cues asked first may unduly influence the character of events reported in a survey.

Another illustration of cuing or example effects comes from the ancestry question in the U.S. census. "English" appeared first in the list of examples following the ancestry question in 1980, but was dropped in 1990. There was a corresponding decrease from 1980 to 1990 of about 17 million persons reporting English ancestry. There were also large increases in the numbers reporting German, Acadian/Cajun, or French-Canadian ancestry, apparently due to the listing of these ancestries as examples in 1990 but not 1980, or their greater prominence in the 1990 list. These effects of examples, and their order, may occur because respondents write in the first ancestry listed that applies to them. In a related question, examples did not have the same effect. Providing examples in the Hispanic origin item increased reporting of specific Hispanic origin groups, both of example groups and of groups not listed as examples, apparently because examples helped

communicate the intent of the question.

Tools for Pretesting and Evaluating Questions

It has always been considered good survey practice to pretest survey questions to ensure they can be administered by interviewers and understood and answered by respondents. Historically, such pretests involved interviewers completing a small number of interviews and being debriefed. Problems were identified based on interview results, such as a large number of "don't know" responses, or on interviewers' reports of their own or respondents' difficulties with the questions. This type of pretest is still valuable, and likely to turn up unanticipated problems. (For automated instruments, it is essential also to test the instrument programming.) However, survey researchers have come to appreciate that many questionnaire problems are likely to go undetected in a conventional pretest, and in recent decades the number and sophistication of pretesting methods have expanded. The new methods have led to greater awareness that survey questions are neither asked nor understood in a uniform way, and revisions based on pretest results appear to lead to improvements. However, questions remain about the validity and reliability of the methods and also the relationship between the problems they identify and measurement errors in surveys. Because the methods appear better able to identify problems than solutions, an iterative approach involving pretesting, revision, and further pretesting is advisable. (A largely unmet need concerns pretesting of translated questionnaires. For cross-national surveys, and increasingly for intranational ones, it is critical to establish that a questionnaire works and produces comparable responses in multiple languages.)

Expert Appraisal and Review

Review of a questionnaire by experts in questionnaire design, cognitive psychology, and/or the relevant subject matter is relatively cost-effective and productive, in terms of problems identified. Nonexpert coders may also conduct a systematic review using the questionnaire appraisal scheme devised by Lessler and Forsyth (see Schwarz and Sudman) to identify and code cognitive problems of comprehension, retrieval, judgment, and response generation. Automated approaches advanced by Graesser and colleagues apply computational linguistics and artificial intelligence to build computer programs that identify interpretive problems with survey questions (see Schwarz and Sudman).

Think-Aloud or Cognitive Interviews

This method was introduced to survey researchers from cognitive psychology, where it was used by Herbert Simon and colleagues to study the cognitive processes involved in problem-solving. The procedure as applied in surveys is to ask laboratory subjects to verbalize their thoughts–to think out loud–as they answer survey questions (or, if the task involves filling out a self-administered questionnaire, to think aloud as they work their way through the questionnaire). Targeted probes also may be administered (e.g., "What period of time are you thinking of here?") Tapes, transcripts, or summaries of respondents' verbal reports are reviewed to reveal both general strategies for answering survey questions and difficulties with particular questions. Cognitive interviews may be concurrent or retrospective, depending on whether respondents are asked to report their thoughts and respond to probes while they answer a question, or after an interview is concluded. Practitioners vary considerably in how they conduct, summarize, and analyze cognitive interviews, and the effects of such procedural differences are being explored.
The verbal reports elicited in cognitive interviews are veridical if they represent information available in working memory at the time a report is verbalized, if the respondent is not asked to explain and interpret his own thought processes, and if the social interaction between cognitive interviewer and subject does not alter a respondent's thought process, according to Willis (see Presser *et al.*). Cognitive interviewing has proved to be a highly useful tool for identifying problems with questions, although research is needed to assess the extent to which problems it identifies translate into

difficulties in the field and errors in data.

Behavior Coding

This method was originally introduced by Cannell and colleagues to evaluate interviewer performance, but has come to be used more frequently to pretest questionnaires. Interviews are monitored (and usually tape recorded), and interviewer behaviors (e.g., "Reads question exactly as worded" and "Reads with major change in question wording, or did not complete question reading") and respondent behaviors (e.g., "Requests clarification" and "Provides inadequate answer") are coded and tabulated for each question. Questions with a rate of problem behaviors above a threshold are regarded as needing revision. Behavior coding is more systematic and reveals many problems missed in conventional pretests. The method does not necessarily reveal the source of a problem, which often requires additional information to diagnose. Nor does it reveal problems that are not manifested in behavior. If respondents and interviewers are both unaware that respondents misinterpret a question, it is unlikely to be identified by behavior coding. Importantly, behavior coding is the only method that permits systematic evaluation of the assumption that interviewers administer questions exactly as worded.

Respondent Debriefing or Special Probes

Respondents may be asked directly how they answered or interpreted specific questions or reacted to other aspects of the interview. Survey participants in effect are asked to assume the role of informant, rather than respondent. Probes to test interpretations of terminology or question intent are the most common form of debriefing question, and their usefulness for detecting misunderstandings is well documented by Belson, Cannell, and others. For example, the following probes were asked following the previously discussed question about doctor visits: "We're interested in who people include as doctors or assistants. When you think of a doctor or assistant, would you include a dentist or not? Would you include a laboratory or X-ray technician or not? ... Did you see any of those kinds of people during the last year?" Specific probes targeted to suspected misunderstandings have proved more fruitful than general probes or questions about respondents' confidence in their answers. (Respondents tend to be overconfident, and there is no consistent evidence of a correlation between confidence and accuracy.) Debriefing questions or special probes have also proved useful for assessing question sensitivity ("Were there any questions in this interview that you felt uncomfortable answering?"), other subjective reactions ("Did you feel bored or impatient?"), question comprehension ("Could you tell me in your own words what that question means to you?"), and unreported or misreported information ("Was there an incident you thought of that you didn't mention during the interview? I don't need details.") Their particular strength is to reveal misunderstandings and misinterpretations of which both respondents and interviewers are unaware.

Vignettes

Vignettes are brief scenarios that describe hypothetical characters or situations. Because they portray hypothetical situations, they offer a less threatening way to explore sensitive subjects. Instead of asking respondents to report directly how they understand a word or complex concept ("What does the term *crime* mean to you?") which has not proved to be generally productive, vignettes pose situations which respondents are asked to judge. For instance:

"I'll describe several incidents that could have happened. We would like to know for each, whether you think it is the kind of crime we are interested in, in this survey.... Jean and her husband got into an argument. He slapped her hard across the face and chipped her tooth. Do you think we would want Jean to mention this incident to us when we asked her about crimes that happened to her?"

The results reveal how respondents interpret the scope of survey concepts (such as crime) as well as the factors influencing their judgments. Research suggests that vignettes provide robust measures of context and question wording effects on respondents' interpretations.

Split-Sample Experiments

Ultimately, the only way to evaluate the effects of variations in question wording, context, etc. on responses is to conduct an experiment in which samples are randomly assigned to receive the different versions. It is essential to ensure that all versions are administered under comparable conditions, and that data are coded and processed in the same way, so that differences between treatments can be unambiguously attributed to the effects of questionnaire variations. Comparison of univariate response distributions shows gross effects, whereas analysis of subgroups reveals conditional or interaction effects. Field experiments can be designed factorially to evaluate the effects of a large number of questionnaire variables on responses, either for research purposes or to select those that produce the best measurements. When a survey is part of a time series and data must be comparable from one survey to the next, this technique can be used to calibrate a new questionnaire to the old.

Conclusion

Survey questionnaire designers aim to develop standardized questions and response options that are understood as intended by respondents and that produce comparable and meaningful responses. In the past, the extent to which these goals were met in practice was rarely assessed. In recent decades, better tools for providing feedback on how well survey questions perform have been introduced or refined, including expert appraisal, cognitive interviewing, behavior coding, respondent debriefing, vignettes, and split-sample experiments. Another advance is new theoretical perspectives that help make sense of the effects of question wording and context. One perspective examines the cognitive tasks in which a respondent must engage to answer a survey question. Another examines the pragmatics of communication in a survey interview. Both have shed light on the response process, although difficult problems remain unsolved. In addition, both perspectives suggest limits on the ability to fully achieve standardization in surveys. New theory and pretesting tools provide a scientific basis for decisions about construction of survey questionnaires.

Further Reading

Belson, W. A. (1981. *The Design and Understanding of Survey Questions.* London: Gower.

Biderman, A. D., Cantor, D., Lynch, J. P., and Martin, E. (1986). *Final Report of the National Crime Survey Redesign Program.* Washington DC: Bureau of Social Science Research.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks CA: Sage Publications.

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

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.

Schaeffer, N. C. and Presser, S. 2003. "The science of asking questions." *Annual Review of Sociology* 29:65-88.

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York: Academic Press.

Schwarz, N. and Sudman, S. (eds.) (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey Bass.

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

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