CHAPTER EIGHT

EVALUATING SURVEY QUESTIONS

8.1 INTRODUCTION

Chapter 2 described two sources of error in survey estimates: errors of nonobservation (when the characteristics of the respondents do not match those of the population from which they were drawn) and observation or measurement error (when the answers to the questions are not good measures of the intended constructs). This chapter examines methods for evaluating survey questions and for determining how much measurement error they introduce into survey statistics.

Question evaluation has two components. First, it assesses the issues discussed in Chapter 7, such as how well questions are understood or how difficult they are to answer, which affect the quality of measurement. Survey researchers evaluate question comprehension, difficulty in memory retrieval, and related issues primarily by observing people trying to understand and answer the questions. The assumption is that questions that are easily understood and that produce few other cognitive problems for the respondents introduce less measurement error than questions that are hard to understand or that are difficult to answer for some other reason. Second, question evaluation assesses how well the answers correspond to what we are trying to measure, that is, directly estimating measurement error. To do this, survey methodologists either compare the survey answers to some other measure or repeat the measurement. The comparison to other measures addresses issues of validity or response bias; the repetition of the measure addresses issues of reliability or response variance. We discuss the practical techniques for estimating these errors in Section 8.9.

There are three distinct standards that all survey questions should meet:

1) Content standards (e.g., are the questions asking about the right things?) — content standards
2) Cognitive standards (e.g., do respondents understand the questions consistently; do they have the information required to answer them; are they willing and able to formulate answers to the questions?) — cognitive standards
3) Usability standards (e.g., can respondents and interviewers, if they are used, complete the questionnaire easily and as they were intended to?) — usability standards

Different evaluation methods are relevant to the three standards. After discussing some of the major tools for assessing questions, we will return to the question of how the different tools provide information related to each standard.

One of the major advances in survey research in the past 20 years has been increased attention to the systematic evaluation of questions. There are at least five different methods that researchers can use to evaluate draft survey questions:

expert review

1) Expert reviews, in which subject matter experts review the questions to assess whether their content is appropriate for measuring the intended concepts, or in which questionnaire design experts assess whether the questions meet the three standards for questions given above.
2) Focus group discussions, in which the survey designers hold a semi-structured ("focused") discussion with members of the target population to explore what they know about the issues that the questionnaire will cover, how they think about those issues, and what terms they use in talking about them.
3) Cognitive interviews, in which interviewers administer draft questions in individual interviews, probe to learn how the respondents understand the questions, and attempt to learn how they formulate their answers.
4) Field pretests, in which interviewers conduct a small number of interviews (typically, fewer than 100) using sampling and field procedures similar to the full-scale survey and in which (a) debriefings with the interviewers may be held (to gain their insights into the problems they had in asking the questions or those the respondents had in answering them); (b) data may be tabulated and reviewed for signs of trouble (such as items with high rates of missing data); and (c) recordings of interviews may be made and behavior coded (to provide quantitative data about questions that are difficult to read as worded or hard for respondents to answer; see Fowler and Cannell, 1996).
5) Randomized or split-ballot experiments, in which different portions of the pretest sample receive different wordings of questions attempting to measure the same thing (Fowler, 2004; Moore, Pascale, Doyle, Chan, and Griffiths, 2004; Schuman and Presser, 1981; Tourangeau, 2004).

The following sections describe how these evaluation activities are carried out in practice.

8.2 EXPERT REVIEWS

As we noted earlier, both subject matter experts and questionnaire design experts may review a draft of the questionnaire. We focus on the role of questionnaire design experts, but emphasize that a review of the questionnaire by substantive experts is also essential to ensure that the questionnaire collects the information needed to meet the analytic objectives of the survey. The experts review the wording of questions, the structure of questions, the response alternatives, the order of questions, instructions to interviewers for administering the questionnaire, and the navigational rules of the questionnaire.

Sometimes, the experts employ checklists of question problems. Over the years, many writers have published lists of principles for good survey questions, starting at least as far back as Payne (1951). Initially, the lists were based primarily on the writers' opinions and judgments. Over time, they have been increasingly informed by cognitive testing, behavior coding of pretest interviews, or psy-

chometric evaluations. We have presented our own list of guidelines (adapted from Sudman and Bradburn, 1982) in Section 7.4.

A difficulty with the items on such lists is that they are subject to interpretation and judgment. Thus, a basic standard for survey questions is that they should be understood in the same way by all respondents, and this understanding should match the one the authors of the question intended. There is no argument about the principle, but even experts looking at the same question can disagree about whether a question is likely to be ambiguous. For that reason, the primary use for a checklist is to guide preliminary review of questions that, in turn, are targeted for some additional form of testing.

There have been several attempts to develop more detailed and explicit systems for assessing potential problems with questions. Lessler and Forsyth (1996) present one of the most detailed, distinguishing more than 25 types of problems with questions, largely derived from a cognitive analysis of the response process similar to the one presented in Section 7.2. Graesser and his colleagues (Graesser, Bommareddy, Swamer, and Golding, 1996) distinguish 12 major problems with questions, most of them involving comprehension issues, and they have developed a computer program that can apply these categories to draft questions, serving as an automated but rough expert appraisal (Graesser, Kennedy, Wiemer-Hastings, and Ottati, 1999). The dozen problems distinguished by Graesser and colleagues include:

1) Complex syntax
2) Working memory overload
3) Vague or ambiguous noun phrase
4) Unfamiliar technical term
5) Vague or imprecise predicate or relative term
6) Misleading or incorrect presupposition
7) Unclear question category
8) Amalgamation of more than one question category
9) Mismatch between the question category and the answer options
10) Difficult to access (that is, recall) information
11) Respondent unlikely to know answer
12) Unclear question purpose

Graesser's model of the question answering process assumes that the listener first decides what type of question has been posed (e.g., a yes–no or a why question); thus, when the type of question is unclear or more than one type is involved, it creates problems for respondents.

8.3 Focus Groups

Before developing a survey instrument, researchers often recruit groups of volunteers to participate in a systematic discussion about the survey topic. A "focus group" discussion is a discussion among a small number (six to ten) of target population members guided by a moderator (see Krueger and Casey, 2000). The group members are encouraged to express their viewpoints and to feel comfortable disagreeing with the perspectives of others. Group members are free to influence one another in their ideas.

focus group

The researcher spends considerable effort structuring the discussion topics in order to target key issues in the subject area. At an early stage of survey development, focus groups might help the researcher learn about how members of the target population understand the concepts in the questionnaire, what terminology they used in discussing them, what common perspectives are taken by the population on key issues, and so on. Groups might discuss reactions to alternative recruitment protocols for the sample members or perceptions of the sponsor of the survey.

The moderator attempts to create an open, relaxed, permissive atmosphere. Good moderators subtly keep the group on target and make seamless transitions across topics. Good moderators encourage all focus group members to participate, drawing out the shy members, and politely closing off the dominant speakers. They listen carefully to each participant's comments, and when new observations are made, seek reactions of other group members to the observations. The moderators are guided by a set of topic areas that are to be discussed. They are not scripted in their questions or probes. Thus, they need to know the research goals quite well in order to answer questions from the group members.

Usually, there is an attempt to choose focus group members who are homogeneous on key dimensions related to the topic. For example, a focus group in preparation for a survey of employment search might separate those persons interested in salaried versus hourly positions. If the survey is to be conducted on diverse subpopulations, then separate focus groups may be mounted for each subpopulation.

Sometimes, focus groups are held in specially designed rooms, with one-way mirrors, to permit the research team to observe the group, and to allow unobtrusive audio/video recording of the group. Sometimes, the product of the group is a set of written notes summarizing key inputs; other times, a full transcript of the group discussion is produced. With videotaped groups, edited videotape segments can summarize the key findings.

Focus groups are common tools in the early stages of questionnaire development, in order to learn what respondents know about the topic of the survey. Focus groups have three main attractions to the questionnaire designer:

1) They are an efficient method for determining what potential respondents know and what they do not know about the survey topics, and how they structure that knowledge. For example, in a survey on health insurance, it is useful to know what types of insurance plans respondents are aware of and what they know (or think they know) about each type of plan. It is also useful to have a sense of which issues or dimensions respondents think are important (and which they think are unimportant). Finally, it is helpful to know how potential respondents think about the issues and how they categorize or group them. For example, in a survey on health insurance, respondents may see health maintenance organizations (HMOs) as very different from other types of health service plans. This sort of information may help the researchers structure the questionnaire to promote the most accurate reporting.
2) Focus groups are also a good method for identifying the terms that respondents use in discussing the topic and exploring how they understand these terms. A key goal in designing survey questions is to use

words that are familiar to the respondents and consistently understood by them. Focus groups provide an excellent opportunity to hear how potential respondents spontaneously describe the issues involved and how they understand candidate phrases or terms.

3) Survey questions ask respondents to describe something they know about. Whether the questions concern subjective states (such as feelings, opinions, or perceptions) or objective circumstances or experiences, the questions are best designed if researchers have a firm grasp of the underlying realities that respondents have to report. Thus, a final function of focus groups is to convey to researchers what respondents have to say on the survey topic. Leading a focus group through the various topics to be covered by a survey and getting a sense of the range of experiences or perceptions that respondents will be drawing upon to form their answers, enables researchers to write questions that fit the circumstances of the respondents.

The strength of the focus group is that it is an efficient way to get feedback from a group of people. There are, however, three main limitations of focus groups:

1) Participants in focus groups are not necessarily representative of the survey population, so one cannot generalize about the distribution of perceptions or experiences from focus groups alone.
2) A focus group is not a good venue for evaluating wording of specific questions or for discovering how respondents arrive at their answers. Focus groups can give a sense of the range and kinds of differences among members of the survey population. However, assessing the wording of specific questions and evaluating the cognitive issues associated with the questions are done more easily with a one-on-one testing protocol.
3) Because the information gathered from the discussion is rarely quantitative, there is the potential for the results and conclusions to be unreliable, hard to replicate, and subject to the judgments of those who are conducting the groups.

Despite these limitations, focus groups are efficient ways of gathering qualitative information about the survey topic from the perspective of the target population prior to imposing the structure of a survey questionnaire.

8.4 COGNITIVE INTERVIEWS

In 1983, the U.S. National Research Council (NRC) held a workshop that brought cognitive psychologists and survey research methodologists together to explore their potential mutual interests. One of the outgrowths of the workshop was that survey researchers began to explore the value of techniques developed by cognitive psychologists to find out how people understand and answer questions (Jabine, Straf, Tanur, and Tourangeau, 1984; Schwarz and Sudman, 1992).

Schuman and Presser (1981) and Belson (1981) had presented evidence indicating considerable misunderstandings of survey questions years before; the papers spawned by the NRC workshop sparked more widespread interest in learning how questions are understood and answered by survey respondents.

cognitive interviewing

protocol analysis

One of the methods discussed at the workshop was the use of cognitive interviewing to test survey questions. Cognitive interviewing is based on a technique called "protocol analysis" that was invented by Simon and his colleagues (see, for example, Ericsson and Simon, 1980, 1984). In a "protocol analysis," subjects think aloud as they work on the problems and their verbalizations are recorded. Simon was interested in the processes by which people solve different kinds of problems, such as proving simple mathematical theorems or playing chess. The term "cognitive interviewing" is used somewhat more broadly to cover a range of cognitively inspired procedures. These include:

1) Concurrent think-alouds (in which respondents verbalize their thoughts while they answer a question)
2) Retrospective think-alouds (in which respondents describe how they arrived at their answers either just after they provide them or at the end of the interview)
3) Confidence ratings (in which respondents assess their confidence in their answers)
4) Paraphrasing (in which respondents restate the question in their own words)
5) Definitions (in which respondents provide definitions for key terms in the questions)
6) Probes (in which respondents answer follow-up questions designed to reveal their response strategies)

This list is adapted from a longer list given by Jobe and Mingay (1989); see also Forsyth and Lessler (1992).

As the list suggests, there is no single way that cognitive interviewing is done. Typically, the respondents are paid volunteers and the interview includes some draft questions along with probes or other procedures for discovering how the respondents understood the questions and arrived at their answers. Respondents may also be asked to think aloud as they answer some or all of the questions. The interviewers may be research scientists, cognitive psychologists, experts in survey question methodology, interviewers with special training in question evaluation, or standardized interviewers with no special training.

Different organizations and different interviewers use different techniques or mixes of techniques to gather information in cognitive tests. Some rely on prescripted probes; others emphasize think-alouds. Some ask for retrospective protocols immediately after a question is administered; others collect them at the end of the interview. There are also different methods for recording the information from cognitive interviews, which range from the formal (e.g., videotaping the interviews and making and transcribing audiotapes) to the informal (having the interviewer make notes during the interviews).

Although the use of cognitive testing is growing and the technique appears to yield valid insights, there is a critical need for empirical studies of the reliability of findings from cognitive interviews, their value in improving data quality,

and the significance of the many variations in the way cognitive interviews are done. DeMaio and Landreth (2004) have conducted the most comprehensive study to date, which showed that different approaches to cognitive testing can produce similar results. Still, their study shows considerable differences among three teams assessing the same questions in which questions they identified as having problems, which problems they thought the questions had, and how they proposed to fix them. In addition, observational studies examining what happens during cognitive interviews indicate considerable variation across interviewers (Beatty, 2004). On the other hand, Fowler (2004) provides examples of questions, revised based on cognitive testing, that result in apparently better data. As yet, though, evidence is limited about the extent to which cognitive testing generally improves survey data (Willis, DeMaio, and Harris-Kojetin, 1999; Forsyth, Rothgeb, and Willis, 2004).

Presser and Blair (1994) on Alternative Pretesting Methods

Presser and Blair compared four pretesting methods.

Study design: Separate pretesting staffs evaluated a common "test" questionnaire of 140 items, in an initial round and a revised round. First, eight telephone interviewers collected 35 first round and 43 second round interviews in a traditional pretest, with a debriefing assessment. Second, the researchers examined behavior coding from the interviews. Third, three cognitive interviewers interviewed a total of about 30 respondents, using follow-up probes and think-aloud techniques. Fourth, two panels of questionnaire experts identified problems in the questionnaire.

Findings: Expert panels identified on average 160 problems compared to about 90 for conventional pretests, cognitive interviews, and behavior coding. Pretests and cognitive interviewing showed great variability over trials. Behavior coding and expert panels were the most reliable in types of problems found. Pretests and behavior coding tended to find interviewer administration problems. Cognitive interviews found comprehension problems and no interviewer problems. The expert panel was the most cost-effective method.

Limitations of the study: Only one questionnaire was used to evaluate the methods. Telephone pretest results may not imply similar findings for face-to-face pretests. There was no distinction made between important and trivial problems. The ability of the researchers to solve the problems was not examined.

Impact of the study: The study led to recommendations to use more expert panels for questionnaire development.

8.5 FIELD PRETESTS AND BEHAVIOR CODING

"Pretests" are small-scale rehearsals of the data collection conducted before the main survey. The purpose of a pretest is to evaluate the survey instrument as well as the data collection and respondent selection procedures. Pretests with small samples (often done by relatively small numbers of interviewers) have been standard practice in survey research for a long time.

Historically, pretests have yielded two main types of information about the survey and the survey questionnaire. First, the views of the interviewers have often been solicited in "interviewer debriefings." These are a bit like focus groups with the pretest interviewers, in which the interviewers present their conclusions about problem questions and other issues that surfaced in the pretest. Often, the interviewers offer suggestions about how to streamline the procedures or improve the questions. The second type of informa-

pretests

interviewer debriefing

tion to emerge from pretests is quantitative information based on the responses. The data collected during a pretest are often entered and tabulated. The survey designers may look for items that have high rates of missing data, out-of-range values, or inconsistencies with with other questions. In addition, items with little variance (that is, items that most respondents answer the same way) may be dropped or rewritten.

behavior coding

Tape recording pretest interviews, then making systematic observations of how questions are read and answered, can also provide useful information about the questions (Oksenberg, Cannell, and Kalton, 1991). "Behavior coding" is the systematic classification and enumeration of interviewer–respondent interaction to describe the observable behaviors of the two persons related to the question-and-answer task. Table 8.1 gives some examples of codes that are used for each question in the questionnaire, for each interview coded.

Using codes like those in Table 8.1, for each interview coded, the behavior coder makes judgments about whether the interviewer reading of the question followed training guidelines and which respondent behaviors were exhibited. The resulting dataset includes the behaviors coded for each question-and-answer sequence for each interview. The question designer then analyzes these data by computing statistics for each question, such as:

1) The percentage of interviews in which it was read exactly as worded.
2) The percentage of interviews in which the respondent asked for clarification of some aspect of the question.

Table 8.1. Examples of Behavior Codes for Interviewer and Respondent Behaviors

Code Category	Description
Interview Questioning Behaviors (choose one)	
	1. Reads question exactly as worded
	2. Reads questions with minor changes
	3. Reads questions so that meaning is altered
Respondent Behaviors (check as many as apply)	
	1. Interrupts question reading
	2. Asks for clarification of question
	3. Gives adequate answer
	4. Gives answer qualified about accuracy
	5. Gives answer inadequate for question
	6. Answers "don't know"
	7. Refuses to answer

3) The percentage of interviews in which the respondent did not initially provide an adequate answer to the question so that the interviewer had to probe or offer an explanation to obtain a codeable answer.

There are various aspects of the question-and-answer process that might be of interest and could be coded. An important area for further research is to identify behaviors that can be reliably coded during interviews and that have implications for the quality of data that are being collected.

Integrating behavior coding into pretesting simply involves tape-recording the interviews, with respondent permission of course, then tabulating the rates at which the behaviors noted above occur. A particular value that behavior coding adds to a standard pretest is that the results are systematic, objective, and replicable. As Fowler and Cannell (1996) report, when two interviewing staffs independently tested the same set of questions, the correlations between the rates of the above behaviors on each question were between 0.75 and 0.90. This indicates that questions consistently and reliably produce high or low rates of these behaviors, regardless of which interviewers are asking the questions.

8.6 RANDOMIZED OR SPLIT-BALLOT EXPERIMENTS

Sometimes, survey designers conduct studies that experimentally compare different methods of data collection, different field procedures, or different versions of the questions. Such experiments can be done as stand-alone studies or as part of a field pretest. When random portions of the sample get different questionnaires or procedures, as is typically the case, the experiment is called a "randomized" or "split-ballot" experiment. Tourangeau (2004) describes some of the design issues for such studies and cites a number of examples of split-ballot experiments. Of our sample surveys, the NSDUH has been especially active in using split-ballot experiments; it has conducted a number of major split-ballot studies examining how reporting of illicit drug use is affected by different modes of data collection and different wording of the questions (see Turner, Lessler, and Devore, 1992; Turner, Lessler, George, Hubbard, and Witt, 1992; and, for a more recent example, Lessler, Caspar, Penne, and Barker, 2000). Similarly, when the Current Population Survey questionnaire required an overhaul of questions about unemployment, a major experiment was carried out to compare the old version of the questions to the new version (Cohany, Polivka, and Rothgeb, 1994). That way, it was clear how much of the change in the monthly unemployment rate (which is derived from the CPS data) was due to the changeover in the questionnaires.

randomized experiments

split-ballot experiments

Experiments like these offer clear evidence of the impact on responses of methodological features—differences in question wording, question order, the mode of data collection, and so on. Unfortunately, although they can demonstrate that the different versions of the instruments or procedures produce different answers, many split-ballot experiments cannot resolve the question of which version produces better data. An exception occurs when the study also collects some external validation data against which the survey responses can be checked. Results are also interpretable when there are strong theoretical reasons for deciding that one version of the questions is better than another. For example, Turner and his colleagues concluded that self-administration improved reporting of illicit

Oksenberg, Cannell, and Kalton (1991) on Probes and Behavior Coding

In 1991, Oksenberg and coworkers reported a study about evaluating questions using behavior coding.

Study design: Six telephone interviewers used a questionnaire of 60 items on health-related topics assembled from questions used on existing surveys. Behavior coding identified some problems with the questions. The researchers revised the questionnaire and took 100 additional interviews.

Findings: The three questions below showed the following behavior coding results:

1) What was the purpose of that visit (to a health care person or organization)?
2) How much did you pay or will you have to pay out of pocket for your most recent visit? Do not include what insurance has paid for or will pay for. If you don't know the exact amount, please give me your best estimate.
3) When was the last time you had a general physical examination or checkup?

Percent of problems per question

Question	1	2	3
Interviewer action			
Slight wording change	2	30	3
Major wording change	3	17	2
Respondent action			
Interruption	0	23	0
Clarification request	2	10	3
Inadequate answer	5	17	87
"Don't know"	0	8	12

The first question was relatively unproblematic. The second caused both the interviewer and respondent conversational problems. The third had the ambiguous phrase, "general physical examination" and the lack of a clear response format. By changing Question 2 so that it did not appear to be completed prematurely, interruptions were reduced.

Limitations of the study: The study did not identify how to fix problems found by behavior coding.

Impact of the study: The study showed how some structural problems with questions could be reliably detected from the question-and-answer behaviors.

drug use because reporting of drug use increased. A number of earlier studies had shown that respondents tend to underreport their drug use, so an increase in reporting is likely to represent an improvement. Fowler (2004) describes several other split-ballot experiments in which, despite the absence of validation data, it seems clear which version of the questions produced the better data.

Fowler's examples involve fairly small samples (some based on fewer than 100 cases), because of the time and expense involved. If detecting relatively small differences between experimental groups is important, large samples of respondents may be needed. For many surveys with modest budgets, adding the cost of even a small split-ballot experiment may seem excessive. For these reasons, split-ballot experiments before surveys are not routine. Nonetheless, they do offer the potential to evaluate the impact of proposed wording changes on the resulting data that other testing approaches do not provide.

8.7 APPLYING QUESTION STANDARDS

The different methods we have discussed vary in the kinds of problems they are best suited to identify. Here we discuss which methods are suitable for evaluating whether the questions meet our three standards for survey questions.

The "content standard" for questions is whether or not they are asking for the right information. This has to be assessed from two perspectives. First, from the point of view of the analysts, the questions must gather the information needed to address the research objectives. The only way to assess this is to ask the experts—the analysts or other subject matter specialists—whether the questions provide the information needed for the analysis. The other key issue is whether the respondents can actually provide this information. Surveys can only provide useful informa-

content standard

tion if respondents can answer the questions with some degree of accuracy. The primary ways to assess how well respondents can answer candidate questions are focus group discussions and cognitive interviews. Through focus groups, we can learn what potential respondents know. From cognitive interviews, we can see whether a particular set of questions can be consistently answered and whether the answers actually provide the information the analysts are seeking.

cognitive standard

"Cognitive standards," whether respondents can understand and answer the questions, are most directly assessed by cognitive testing. That is what cognitive interviews are designed to do. However, there are three other question evaluation activities that can make a contribution to identifying cognitive problems with questions:

1) Focus group discussions can identify words that are not consistently understood, concepts that are ambiguous and questions that respondents are unable to answer
2) Expert reviews often can flag ambiguous terms and concepts, as well as response tasks that are difficult to perform, prior to any cognitive testing
3) Behavior coding of pretest interviews can identify questions that are unclear or that respondents have trouble answering

usability

The assessment of "usability," how well a survey instrument can be used in practice, is the primary aim of a field pretest. In addition, prior to a pretest, expert reviews of the questions can identify questions that are likely to pose problems for respondents or interviewers. Usability testing is most valuable in self-administered questionnaires. In controlled laboratory conditions or in typical survey settings, survey staff can observe respondents handling the survey materials, attempting to understand the task, and performing the task. With computer-assisted instruments (Couper, Hansen, and Sadowsky, 1997; Hansen and Couper 2004; Tarnai and Moore, 2004), the computer itself might be used to time keystrokes, to measure the extent of backward movement in the questionnaire, and to compute the rate of illegal entries. Although the laboratory may not turn up all the problems likely to crop up in the field, the problems identified there are likely to be even worse under realistic data collection conditions.

8.8 SUMMARY OF QUESTION EVALUATION TOOLS

All of the techniques for evaluating survey questions discussed in this chapter have potential contributions to make and all of them have limitations. Here we summarize some of these virtues and limitations.

1) Expert content review provides the important perspective of what the users of the data need in order to meet the analytic objectives of the survey. However, it provides no information about what the best questions are, the ones that respondents can answer most accurately in order to provide the necessary information.
2) Systematic review of the questions by questionnaire design experts is perhaps the least expensive method and the easiest to carry out (Presser

and Blair, 1994; see box on page 265). Still, it is only as good as the experts. Experts may disagree about whether the questions are clear or the response tasks they pose are too difficult for respondents to carry out. As a result, techniques such as cognitive testing with real potential respondents are needed to evaluate the questions empirically. On the other hand, a growing number of characteristics of questions have been identified as consistently posing problems. To the extent that it is possible to expand the list of question characteristics that are likely to cause problems, the value of expert review of the questions can be proportionately increased.

3) Focus group discussions provide an efficient way to get the ideas, perceptions, and contributions of six to ten people at a time about issues of practical relevance to designing the questions. However, because it is a group setting, focus groups are not the best method for investigating how individuals understand the specific questions or how they go about answering them.
4) Cognitive testing is a useful method to find out how individuals understand questions and arrive at their answers to them. However, cognitive testing usually involves a small number of people, who are not necessarily representative of an entire target population. Thus, a major concern about cognitive testing is the generalizability of the results. We cannot know how the distribution of problems or issues found in a group of cognitive interview respondents will generalize to the target population. We also need to be concerned that paid volunteers under laboratory conditions may be able and willing to perform tasks that respondents under realistic survey conditions will not. Finally, there is the danger that different cognitive interviewers may produce different conclusions, perhaps even steering the interviews to produce evidence of problems (Beatty, 2004). A related issue is that cognitive interviews yield unsystematic impressions about the problems with the questions rather than objective data (Conrad and Blair, 1996).
5) Usability testing in laboratory settings has strengths and limitations that parallel those of cognitive testing.
6) Field pretests are the best way to find out how instruments and field procedures work under realistic conditions. Through behavior coding, researchers can gain systematic information about how the question-and-answer process in fact is performed under real conditions. It is generally useful to tabulate the data and to debrief the pretest interviewers. The limitations of field pretests are that, because researchers are trying to replicate realistic survey procedures, there is not great flexibility to probe and understand the nature of the problems that interviewers and respondents face.

One critical question is the extent to which the various techniques produce the same information. Table 8.2 summarizes the results of several studies that compare multiple methods of item evaluation. The first of these studies, done by Presser and Blair (1994) compared the results from conventional pretests, expert reviews, cognitive testing, and behavior coding of a pretest (see box on page 265). They found that there was some overlap in the problems found, but the approaches did not yield the same results. The expert reviews and cognitive inter-

Table 8.2. Studies Comparing Question Evaluation Methods

Presser and Blair (1994)		
Methods Tested	**Criteria**	**Conclusions**
1. Coventional pretest 2. Behavior coding 3. Cognitive interviews 4. Expert panels	• Number of problems found • Type of problem detected (problems were classified into four categories) • Consistency across trials with the same method	1. Conventional pretests and behavior coding found the most interviewer problems. 2. Expert panels and cognitive interviews found the most analysis problems. 3. Expert panels and behavior coding were more consistent across trials and found more types of problems. 4. Behavior coding was most reliable but provided no information about the cause of a problem, did not find analysis problems, or distinguish between respondent-semantic and respondent-task problems. 5. Expert panels were most cost-effective. 6. Most common problems were respondent-semantic.
Willis, Schechter, and Whitaker (1999)		
Methods Tested	**Criteria**	**Conclusions**
1. Cognitive interviewing (done by interviewers at two organizations) 2. Expert review 3. Behavior coding	• Number of problems found • Consistency within and across methods regarding the presence of a problem (measured by the correlation across methods and organizations between the percent of the time items were classified as having a problem) • Type of problems found (based on a five-category coding scheme)	1. Expert review found the most problems. 2. The correlation between behavior coding trials was highest (0.79), followed closely by the correlation between the cognitive interviews done by two organizations (0.68). 3. Across methods of pretesting and organizations, most problems were coded as comprehension/communication; there was a high rate of agreement in the use of subcodes within this category across techniques.
Rothgeb, Willis, and Forsyth (2001)		
Methods Tested	**Criteria**	**Conclusions**
Three research organizations tested three questionnaires, using each of these methods and coding problems according to a classification scheme developed by the authors: 1. Informal expert review 2. Formal cognitive appraisal 3. Cognitive interviewing	• Number of problems found • Agreement across methods based on summary score for each item (summary scores ranged from 0 to 9 based on whether the item was flagged as a problem item by each technique and each organization)	1. Formal cognitive appraisal (QAS) found the most problems but encouraged a low threshold for problem identification. 2. Informal expert review and cognitive interviewing found similar numbers of problems, but found different items to be problematic. 3. Results across organizations were more similar than across techniques: Moderate agreement across organizations in summary scores (r's range from 0.34 to 0.38). 4. Communication and comprehension problems were identified most often by all three techniques.

(*continued*)

Table 8.2. Studies Comparing Question Evaluation Methods (continued)

Forsyth, Rothgeb, and Willis (2004) (Note: This study is a follow-up to Rothgeb et al., 2001)		
Methods Tested	**Criteria**	**Conclusions**
1. Informal expert review 2. Formal cognitive appraisal (QAS) 3. Cognitive interviewing	• Conducted randomized experiment in a random-digit-dial telephone survey that compared control questionnaire (with the original items pretested in 2001 study) and experimental questionnaire (with revised items designed to fix problems found in the pretest) • Classified items as low, moderate, or high in respondent and interviewer problems, based on behavior coding data and interviewer ratings	1. Items classified as high in interviewer problems during pretesting also had many problems in the field (according to behavior coding and interviewer ratings). 2. Items classified as high in respondent problems during pretesting also had many problems in problems in the field. 3. Items classified as having recall and sensitivity problems during pretesting had higher nonresponse rates in the field. 4. The revised items in the experimental questionnaire produced nonsignificant reductions in item nonresponse and problems found via behavior coding, but a significant reduction in respondent problems (as rated by the interviewers). However, interviewers rated revised items as having more interviewer problems.
DeMaio and Landreth (2004)		
Methods Tested	**Criteria**	**Conclusions**
1. Three cognitive interview methods (three different "packages" of procedures carried out by three teams of researchers at three different organizations) 2. Expert review	• Number of problems identified • Type of problem identified • Technique that identified the problem • Frequency of agreement between organizations/methods	1. The different methods of cognitive interviewing identified different numbers and types of problems. 2. Cognitive interviewing teams found fewer problem questions than did expert reviews, although all three organizations identified as problematic most of the "defective" questions (those for which at least two experts agreed there was a problem of a specific type). 3. The problems identified by the cognitive interviewing teams were also generally found by the experts. 4. Different teams used different types of probes. 5. Upon revising the questionnaires and readministering cognitive interviews, it was found that only one team's questionnaire had fewer problems than the original.

views tended to find more comprehension problems, whereas the pretest found more problems in administering the items (that is, usability issues). The experts flagged the most problems, but not always important problems.

The conclusions of the Presser and Blair study have stood up well over time. For example, a more recent study by Forsyth, Rothgeb, and Willis (2004) took a more critical look at the importance of the problems identified by the various methods and found (as Presser and Blair did) that the techniques find some of the same problems but some unique question problems as well. Experts find the most "problems," but some of those probably do not affect data quality. Weak measures of data quality, a chronic challenge for studies of question design and evaluation, made the results of these evaluation studies inconclusive; we cannot say whether the "problems" really reduce the validity of the answers.

There seems to be little doubt that the different techniques complement each other; each has some obvious strengths and weaknesses compared to the others and they provide information related to different issues. As a result, many surveys use a combination of methods to pretest survey questions. Which specific methods are used often depends on the survey budget and the level of prior experience with the questions. A survey with a new questionnaire but a limited pretesting budget might conduct a few focus groups, an expert review, a round or two of cognitive interviews, and a small field pretest. Expert reviews and cognitive testing can be done cheaply, and both are likely to uncover lots of potential problems (Presser and Blair, 1994; Forsyth, Rothgeb, and Willis, 2004). The focus groups will help align the concepts and terminology used in the questionnaire with those of the respondents. A small field test can detect any operational problems. If an existing questionnaire is being used with minor modifications, the focus groups and field tests might be dropped, and cognitive testing would focus on the new items.

At the other extreme, a large survey is rarely fielded without a correspondingly large field pretest. The risk of a major operational failure is too great to go into the field without substantial pretesting. Census 2000, for example, underwent a series of field tests culminating in a dress rehearsal conducted in 1998 that involved three areas (Sacramento, California; Menominee County, Wisconsin; and an 11 county area including Columbia, South Carolina) and several hundred thousand respondents. The pretesting program that led to the latest NSDUH questionnaire involved several large-scale split-ballot experiments (e.g., Lessler, Caspar, Penne, and Barker, 2000).

However, the major problem with all of these techniques is that they tell us very little, if anything, about how the problems identified affect data quality. There is some evidence that the problems identified by these testing techniques can have major effects on survey estimates. For instance, Fowler (1992) has shown that if key terms in questions are not understood consistently, systematic biases are likely to result. Mangione and his colleagues have also shown that if questions are hard for respondents, so that interviewers have to probe extensively to get adequate answers, it increases the likelihood of interviewer effects and results in with inflated standard errors (Mangione, Fowler and Louis, 1992). Nonetheless, the techniques discussed so far, with the possible exception of split-ballot tests, do not tell us what kind and how much error particular question problems are likely to produce. Moreover, we lack assessments of which testing techniques, either alone or in combination, are best at finding problems that affect survey results. To get estimates on those issues, different kinds of studies and analyses are needed. That is the topic of the next section.

8.9 Linking Concepts of Measurement Quality to Statistical Estimates

Unfortunately, the terminology used within survey methodology for the quality of measurement is not standardized. The two traditions that provide the common terms are psychometrics and sampling statistics. The first focuses on answers to questions by an individual respondent and uses the terms "validity" and "reliability." The second focuses on statistics summarizing all the individual answers and uses the terms "bias" and "variance."

8.9.1 Validity

validity

"Validity" is a term used in somewhat different senses by different disciplines and, even within survey research, it seems to mean different things to different researchers. A common definition of "validity" is the extent to which the survey measure accurately reflects the intended construct; this definition applies in different ways to different items. Unfortunately, this definition does not suggest a specific method of evaluating validity. An early meaning (Lord and Novick, 1968) was based on a simple conceptual model of the measurement process as being one realization of a set of conceptually infinite trials. That is, each survey measurement could (in concept only) be repeated so that each answer from a given respondent to a given question administration is just one trial within that infinite set of trials. Just as we presented in Section 2.3, let

μ_i = the true value of the construct for the ith respondent
Y_{it} = the response to the measure by the ith respondent on the tth trial
ε_{it} = the deviation from the true value of the construct related to the response, Y_i, on the tth trial

Then the conceptual model of the response process is the following. When the question about the construct, μ, is given to the ith respondent in a survey (called the tth trial), instead of providing the answer, μ_i, the respondent provides the answer, Y_{it},

$$Y_{it} = \mu_i + \varepsilon_{it}$$

"Validity" is measured by the correlation (over persons and trials) of two of the terms in the equation above. It is usually defined as the correlation between Y_i and μ_i. This means that a measure is higher in validity when the values of Y_i are on average, closer to those of μ:

$$\text{Validity}(Y) = \frac{\sum_{i,t}(Y_{it}-\bar{Y})(\mu_i-\bar{\mu})}{\sqrt{\sum_{i,t}(Y_{it}-\bar{Y})^2 \sum_{i}(\mu_i-\bar{\mu})^2}} = \text{Correlation}(Y_i, \mu_i)$$

Given this, validities are expressed by numbers that lie between 0.0 to 1.0; the higher the number, the higher the validity. We will describe two ways of estimating validity in practice: using data external to the survey and using multiple indicators of the same construct in one survey.

Estimating Validity with Data External to the Survey. Consider an item we discussed in the previous chapter:

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

The item is a factual item and, at least in principle, the quality of the answer can be determined with reference to a reasonably well-defined set of facts. The respondent has, in fact, made a certain number of eligible medical visits within the period specified and, although there may be some ambiguity as to whether specific episodes should be included (for instance, should respondents count a consultation with a doctor by telephone?), in principle, these ambiguities could be resolved. For this question, if the respondent had two visits in the past 12 months, μ_i is 2. If Y_i is 2, there is no response deviation from the true value for that trial. If, across all respondents, answers similarly agreed with their true number of visits and this occurred in all possible trials, there would be validity of 1.0.

With some survey items, records or other external data might allow us to assess the quality of survey responses. By asserting that whatever value lies in the record for the ith respondent is μ_i, the true value for the respondent can be calculated. If we can further assert that the survey we conduct is one representative of all possible trials, we can compute validity by

$$\text{Estimated validity from one trial} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{\mu}_i - \bar{\hat{\mu}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{\mu}_i - \bar{\hat{\mu}})^2}}$$

where

μ_i = the value of the record variable for the construct, sometimes called the "gold standard" for the study
$\bar{\mu}$ = the mean of the record values across all respondents

This is merely the correlation between the respondent answers and their record values. If the correlation is near 1.0, then the measure is said to have high validity.

Estimating Validity with Multiple Indicators of the Same Construct. A second item is attitudinal, and it is not clear what facts are relevant to deciding whether respondents have answered accurately.

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

This item is intended to assess economic expectations, and the actual state of the economy twelve months later is not directly relevant to the accuracy of the answer. The set of facts relevant to deciding whether a respondent's answer to this question is a good one is not well defined, and these facts are, in any case, subjective ones. Most respondents do not have some preexisting judgment about the future state of the economy but probably consult various beliefs about the economic trends in answering the question. For instance, respondents may consider their beliefs about the trend in the unemployment rate or about the direction of the stock market. It would be very difficult to say which beliefs are relevant to judging the accuracy of the respondents' answers and we do not have direct access to those beliefs anyway. In the notation above, it is not clear what μ_i is for this question for any respondent. Issues of measuring validity are thus more complex.

Correlating survey answers to some "gold standard" derived from another source is ideal. However, it is unusual in survey research to have accurate external information with which to evaluate answers. Moreover, with surveys that ask about subjective states, such as knowledge, feelings, or opinions, there is no possibility of a "gold standard" independent of the respondent's reports. In the absence of some outside standard, evaluation of the validity of most answers rests on one of three kinds of analyses:

1) Correlation of the answers with answers to other survey questions with which, in theory, they ought to be highly related
2) Comparison between groups whose answers ought to differ if the answers are measuring the intended construct
3) Comparisons of the answers from comparable samples of respondents to alternative question wordings or protocols for data collection (split-ballot studies)

The first approach is probably the most common for assessing validity. For example, if an answer is supposed to measure how healthy a person is, people who rate themselves at the high end of "healthy" should also report that they feel better, that they miss fewer days of work, that they can do more things, and that they have fewer health conditions than those who rate themselves at the low end of the scale. The results of such analyses are called assessments of construct validity (Cronbach and Meehl, 1955). If researchers did not find the predicted relationships, it would cast doubt on the validity of the health measure.

For example, the Mental Health Inventory Five Item Questionnaire (MHI-5) is a widely used series of questions designed to measure current psychological well-being (Stewart, Ware, Sherbourne, and Wells, 1992):

> These questions are about how you feel and how things have been going during the last four weeks. For each question, please

give the one answer that comes closest to the way that you have been feeling. How much of the time during the past four weeks:

a) Have you been a happy person?
b) Have you felt downhearted and blue?
c) Have you been a very nervous person?
d) Have you felt calm and peaceful?
e) Have you felt so down in the dumps that nothing could cheer you up?

The response categories include "all of the time," "most of the time," "a good bit of the time," "some of the time," "a little bit of the time," and "none of the time." All of these questions ask people to describe their psychological well-being. An index is created by assigning a number (for example, from 1 to 6) to each of the six possible answers, reverse scoring the negatively worded items so 6 is always a positive response, and adding the answers across the five questions (producing a score that could range from 5 to 30).

We can evaluate the validity of the MHI-5, to see the extent to which it was measuring what it was intended to measure, by looking at the correlation of the total score with other measures. Other indicators of mental health are used as criteria. This is sometimes called "concurrent validity," because it is based on relationships among attributes measured at the same time. Stewart, Ware, Sherbourne, and Wells (1992) find that the MHI-5 correlates –0.94 with a measure of psychological distress, –0.92 with a measure of depression, –0.86 with a measure of anxiety, +0.88 with a measure of positive affect, +0.69 with a measure of perceived cognitive functioning, and +0.66 with a measure of feelings of belonging. The authors conclude that these patterns of association are in the direction and of the order of magnitude one would expect to find with a summary measure of mental health, and, hence, that there is good evidence for the validity of the measure.

Validity assessment can also be applied using sophisticated modeling approaches that simultaneously evaluate the evidence for validity by looking at the patterns and strength of the correlations across numerous measures (e.g., Andrews, 1984; Saris and Andrews, 1991). Consider the MHI-5 items given earlier. The structural modeling approach depicts answers to each of the five items as reflecting the same underlying construct μ_i at different levels of validity λ_α:

$$Y_{\alpha i} = \lambda_\alpha \mu_i + \varepsilon_{\alpha i}$$

Notice that this model is just another small variant of the base error model of $Y_i = \mu_i + \varepsilon_i$. Here, instead of a response to just one item, the equation describes the responses to many items, each subscripted with a different α. The response to each item is viewed as a function of the underlying construct μ_i, which is described by the coefficient λ_α.

For example, a simple model of the measurement process for the MHI-5 can be presented as a path diagram, as in Figure 8.1. The circle at the top represents the underlying construct μ_i. The arrows emanating out of the circle imply that the construct "causes" the values of the indicators ($Y_{\alpha i}$) as a function of the λ_α, $\alpha = 1$,

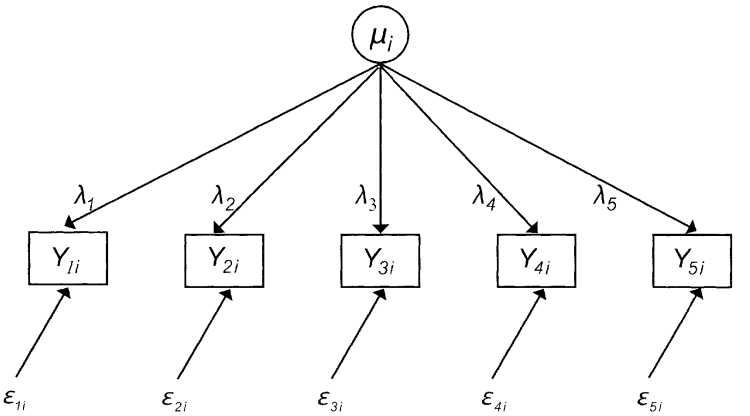


Figure 8.1 Path diagram representing $Y_{\alpha i} = \lambda_\alpha \mu_i + \varepsilon_{\alpha i}$, a measurement model for μ_i.

2, 3, 4, or 5, which are displayed along the arrows. The individual items are presented in squares, one for each of the five items.

Thus, in a more compact form the figure describes five equations:

$$
\begin{aligned}
Y_{1i} &= \lambda_1 \mu_i + \varepsilon_{1i} \\
Y_{2i} &= \lambda_2 \mu_i + \varepsilon_{2i} \\
Y_{3i} &= \lambda_3 \mu_i + \varepsilon_{3i} \\
Y_{4i} &= \lambda_4 \mu_i + \varepsilon_{4i} \\
Y_{5i} &= \lambda_5 \mu_i + \varepsilon_{5i}
\end{aligned}
$$

The λ_α are the validity measures for each of the five items. They can be estimated using a variety of techniques (see Andrews, 1984; Saris and Andrews, 1991; Saris and Gallhofer, 2007).

Another approach to assessing validity is to compare the answers of groups of respondents who are expected to differ on the underlying construct of interest. For example, in the United States, we expect Republicans to be more conservative than Democrats. Thus, we would expect registered Republicans to score higher, on average, than registered Democrats on an item designed to measure conservatism. Such evaluations depend critically on our theory about the relationships among certain variables. If, in the above example, registered Republicans were not different from Democrats, it could mean either that the question was a poor measure of conservatism or that Republicans and Democrats do not, in fact, differ on conservatism. In general, it is difficult or impossible to distinguish poor measurement from inaccurate theories in the assessment of construct validity.

These results can only be interpreted in the context of a good theory. If there is no basis for deciding the likely direction of the errors, then we cannot say which results are more accurate or valid when two data collection protocols produce different results.

8.9.2 Response Bias

The most common confusion in terminology about errors associated with questions is how "validity" relates to "bias." As we pointed out earlier, validity is a function of the correlation between the response and the true value. Thus, validity is a property of individual answers to questions. What happens if there is a systematic deviation in responses away from a true value? For example, Sections 5.3.5 and 7.3.7 note consistent underreporting of socially undesirable traits. In some circumstances, systematic underreporting may not lower the correlation between the responses and their true values. For example, if all respondents underreport their weight by five pounds, the correlation between their reported weight and their true weight is 1.0. However, the mean weight of the respondents would be exactly five pounds less than the mean of the true values. The term "bias" is most often used to describe the impact of systematic reporting errors on summary statistics like sample means.

response bias

The split-ballot approach is designed to test for the presence of bias. Sudman and Bradburn (1982) provide an example. They compared two questions about alcohol consumption:

a) On days when you drink alcohol, how many drinks do you usually have—would you say one, two or three, or more?
b) On days when you drink alcohol, how many drinks do you usually have—would you say one or two, three or four, five or six, or seven or more?

The respondents were randomly assigned to answer one of these two questions. Sudman and Bradburn found that those who were asked question (b) were much more likely than those who were asked question (a) to report having more than three drinks on days that they drank alcohol. The researchers were confident that respondents tended to underreport the amount of alcohol they consumed. Given that premise, they concluded that the answers to the second question were more valid than the answers to the first question. Note that in a fashion similar to that above regarding measurement of validity, this measurement of bias forces the researcher to make some assumption about truth. (When some record system is used to compare survey responses, the assumption is that the records are measured without error.)

To illustrate "bias" in survey statistics, we can start with the same measurement model:

$$Y_{it} = \mu_i + \varepsilon_{it}$$

That is, the response to the question departs from the true value by an error. If the error term has a systematic component, its expected value will not be zero. Bias is introduced when the expected value of the survey observations (the Y's) differs from that of the true scores; that is, there is a systematic deviation across all trials and persons between their response and their true value:

$$\text{Bias}(Y_{it}) = \sum_t \left[\sum_i (Y_{it}) - \mu_i \right]$$

This expression is related to the summary statistic, the mean. The average or expected value of an individual response over all persons (the E_i part of the expression above) is merely the mean of the responses.

Thus, this is the same as saying that the mean of the responses is a biased estimate of the mean of the true values:

$$\text{Bias}(\bar{Y}) = \sum_t \left(\frac{\sum_i Y_{it}}{N} \right) - \frac{\sum_i \mu_i}{N}$$

The first term in the expression above is merely the expected value of the response mean over all trials.

The expression of the bias depends on values of the μ_i, but the expression for the validity above just depends on correlation of the μ_i's with the Y_i's. The notion of bias depends on the existence of a true value. Whether "true values" exist for subjective states, such as knowledge, opinions, and feelings, is controversial. Although psychometricians have sometimes tried to define true scores for subjective constructs like attitudes, as we pointed out, it generally is impossible to get a measure of a true score based on an external source. That is, with subjective constructs, we can only compare two or more reports from the respondent. For this reason, the concept of bias technically only applies to measures of objectively verifiable facts or events.

There are two ways, in practice, that response bias is estimated empirically: using data on individual target population elements from sources external to the survey and using population statistics not subject to the survey measurement error.

Using Data on Individual Target Population Elements. When the existence of a true value is plausible, the practical way to estimate response bias of answers to survey questions is to compare them with some external indicator of the true value. For example, some studies of response bias in surveys have used medical records as the indicators of the true values (Cannell, Marquis, and Laurent, 1977; Edwards, Winn, and Collins, 1996). Cannell and his colleagues compared respondents' reports about whether they had been hospitalized in a specific period with hospital records for that same period. Similarly, Edwards and his associates evaluated the quality of reporting of health conditions by comparing survey reports with the conditions recorded in medical records (Edwards et al., 1994). When records exist with which to compare survey answers, researchers are in the best position to evaluate the quality of reporting.

record check study

Let us examine one of these "record check" studies in more detail. In the study by Cannell and colleagues (Cannell, Marquis, and Laurent, 1977), interviews were conducted in households that included someone who had been hospitalized during the year prior to the interview. The households had been selected using the hospital records, and the researchers determined whether the hospitalizations in the records were in fact reported in the health interview. Table 8.3 shows the percentage of known hospitalizations reported by the length of stay (how many days the patient was in the hospital) and by how many weeks prior to the interview the hospitalization occurred. On average, respondents reported only about 85% of their hospitalizations. However, their likelihood of reporting a hospitalization was greatly affected by the length of stay and the recency of the hos-

Table 8.3. Percentage of Known Hospitalizations Not Reported, by Length of Stay and Time Since Discharge

	Length of Stay		
Time Since Discharge	1 day	2–4 days	5 days
1–20 weeks	21%	5%	5%
21–40 weeks	27%	11%	7%
41–52 weeks	32%	34%	22%

Source: Cannell et al., 1977.

pitalization. More recent hospitalizations were reported better than those that occurred earlier; hospitalizations that involved stays of five or more days were reported better than those that were shorter.

This table is a classic demonstration of how memory deteriorates over time and how more major events are easier to remember than those that are less significant (cf. Figure 7.1). Table 8.3 implies a bias in the estimate of a summary statistic—the mean number of hospitalizations in a group of respondents. The bias of that mean would be relatively greater for the mean number of hospitalizations 41–52 weeks before the interview than for the mean of hospitalizations 1–20 weeks before the interview.

Using Population Statistics Not Subject to Survey Response Error. Sometimes, survey data also can be evaluated in the aggregate, even when the accuracy of the individual reports cannot be determined. For example, a 1975 survey studied gambling behavior (Kallick-Kaufman, 1979). The survey included questions about the amount of money respondents had wagered legally at horse racing tracks. The total amount wagered at horse racing tracks is published. The researchers could not evaluate the quality of a respondent's answer directly. They could compare the overall survey estimate to the published total wagers. This allowed them to estimate net bias of the estimate. They found that the estimates from the survey and those from the racetrack records were remarkably similar. As a result, the researchers concluded that respondents were not, on average, under- or overreporting the amount they wagered at racetracks (Kallick-Kaufman, 1979). Similarly, voter behavior is published the day of elections and survey statistics summarizing voter behavior can be compared to the published results. Without the individual results, however, only the bias of summary statistics that are available publicly can be estimated.

8.9.3 Reliability and Simple Response Variance

reliability

"Reliability" is a measurement of variability of answers over repeated conceptual trials. Reliability addresses the question of whether respondents are consistent or

stable in their answers. Hence, it is defined in terms of a variance component, the variability of ε_{it} over all respondents and trials. In notation,

$$\text{Reliability}\,(Y_{it}) = \frac{\underset{i}{E}(\mu_i - \bar{\mu})^2}{\underset{i}{E}(\mu_i - \bar{\mu})^2 + \underset{i\,t}{E}(\varepsilon_{it} - \bar{\varepsilon})^2} = \frac{\text{Variance of true values}}{\text{Variance of reported values}}$$

If the variance of the response deviations

$$\underset{i,t}{E}(\varepsilon_{it} - \bar{\varepsilon})^2$$

is low, this reliability ratio approaches 1.0 and the measure on the population is said to have "high reliability." If the variability in answers over trials is high (thus producing large response deviations variance), then reliability approaches 0.0.

simple response variance

The survey statistics field does not tend to use the term "reliability," but instead uses the phrase, "simple response variance." It is appropriate to think of "simple response variance" as the opposite of "reliability." When an item has high reliability for a population, then it has low simple response variance. (The term "simple response variance" will be constrasted with "correlated response variance," discussed in Section 9.3).

"Reliability" refers to the consistency of measurement either across occasions or across items designed to measure the same construct. Accordingly, there are two main methods that survey researchers use to assess the reliability of reporting: repeated interviews with the same respondent and use of multiple indicators of the same construct.

reinterview

Repeated Interviews with the Same Respondent. These are sometimes called "reinterview studies." These repeated measures can be used to assess simple response variance with the following assumptions:

1) There are no changes in the underlying construct (i.e., μ_i does not change) between the two interviews.
2) All the important aspects of the measurement protocol remain the same (i.e., these are sometimes referred to as "essential survey conditions").
3) There is no impact of the first measurement on the second responses (i.e., there are no memory effects; the second measure is independent of the first).

In practice, there are complications with each of the three assumptions. However, reinterview studies, in which survey respondents answer the same questions in a second interview as they did in a prior interview, are a common approach to obtaining estimates of simple response variance (e.g., Forsman and Schreiner, 1991; O'Muircheartaigh, 1991). When the survey measurement is repeated, the researcher essentially has two responses for each respondent, Y_{i1} and Y_{i2}.

Using reinterview studies, there are several different statistics commonly calculated to measure the consistency of response over trials:

1) Reliability, as defined above
2) The index of inconsistency, which equals (1 – reliability) — index of inconsistency
3) The simple response variance, which is merely

$$\frac{1}{2N}\sum_i (Y_{i1} - Y_{i2})^2$$

where Y_{i1} and Y_{i2} are the answers in the interview and reinterview, respectively.

4) The gross difference rate, which for a dichotomous variable is merely twice the simple response variance — gross difference rate

None of these measures have distinct advantages over the others, as they are all arithmetic functions of one another. Different survey organizations tend to use different measures of consistency in reporting.

To provide an example of practical uses of such response error statistics, Table 8.4 shows indices of inconsistency for the NCVS (see Graham, 1984). For example, the index of inconsistency for the proportion of the respondents who reported a broken lock or window is 0.146. This corresponds to a reliability coefficient of (1 – 0.146) = 0.854, which is often considered a high level of reliabil-

Table 8.4. Indexes of Inconsistency for Various Victimization Incident Characteristics, NCVS

Question and Category	Index of Inconsistency: Point Estimate	Index of Inconsistency: 95% Confidence Interval
6c. Any evidence offender(s) forced way in building? (multiple response item)		
Broken lock or window	0.146	0.094–0.228
Forced door or window	0.206	0.143–0.299
Slashed screen	0.274	0.164–0.457
Other	0.408	0.287–0.581
13f. What was taken? (multiple-response item)		
Only cash taken	0.276	0.194–0.392
Purse	0.341	0.216–0.537
Wallet	0.189	0.115–0.310
Car	0.200	0.127–0.315
Part of car	0.145	0.110–0.191
Other	0.117	0.089–0.153
7b. Did the person(s) hit you, knock you down, or actually attack you in any way? (single-response item)		
Yes	0.041	0.016–0.108
No	0.041	0.016–0.108

Source: Graham, 1984, Tables 59–60.

ity for survey responses. Note that the indices of inconsistency for the visible evidence and items taken are higher (lower reliability), on average, than the indices for whether the offender attacked the respondent (index of inconsistency = 0.041; reliability = 0.959). This reflects the fact that some types of attributes generated more consistent answers over trials than others.

Using Multiple Indicators of the Same Construct. Another approach to assessing reliability is to ask multiple questions assessing the same underlying construct. This approach is used most often to measure subjective states. This approach makes the following assumptions:

1) All questions are indicators of the same construct (i.e., their expected values are the same).
2) All questions have the same expected response deviations (i.e., their simple response variance or reliability is constant).
3) The measures of the items are independent (i.e., the answer from one item does not influence how the respondent replies to another).

Cronbach's alpha

Cronbach's alpha is a widely used measure of the reliability of such multi-item indices (Cronbach, 1951).

Section 8.9.1 described the five-question mental health index called the MHI-5 and the use of correlations of the index with other mental health indicators as a way of estimating its validity. The reliability of the index generated by combining answers to the items, as measured by Cronbach's alpha (α), depends on the number of items k and their average intercorrelation $\bar{r}$:

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$

For instance, suppose the correlations of each item with all the other items were as given in Table 8.5. The average of the ten correlations is 0.539, so that α is 0.85:

Table 8.5. Illustrative Intercorrelations among MHI-5 Items

Question	Happy Person	Down-hearted, Blue	Nervous	Calm	Down in Dumps
Happy Person	–				
Downhearted, Blue	0.55	–			
Nervous	0.45	0.59	–		
Calm	0.62	0.51	0.54	–	
Down in Dumps	0.49	0.63	0.56	0.45	–

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$
$$= \frac{5(0.539)}{1+4(0.539)} = 0.854$$

A high value of Cronbach's alpha implies high reliability or low response variance. Unfortunately, it can also indicate that the answers to one item affected the responses to another to induce high positive correlation. A low value can indicate low reliability or can indicate that the items do not really measure the same construct.

The MHI-5 rests on the assumption that answers to each of the five questions reflect at least in part the same underlying construct. The answers reflect both the common construct and some item-specific variance because, for example, while "happy" and "calm" both are good emotional states, they are not identical. The evidence that the items reflect a common construct is the high level of intercorrelations among the five items.

Not all multi-item measures derive their validity in this way. An investigator can define a complex concept and then sample different examples of aspects of it that need not be related in any way. An example of this sort of multi-item scale comes from the Consumer Assessment of Health Plans (CAHPS) surveys, which are designed to measure patient experiences in getting health care. The CAHPS surveys use a number of multi-item composite measures designed to summarize patient reports of their experiences; the results are provided to people who are choosing health plans. One example is a composite called Getting Care Quickly, which consists of four questions asking patients how often they (1) got the help they needed when they called a doctor's office for help or advice, (2) got appointments as soon as they wanted them for routine care, (3) saw a doctor as soon as they wanted when they needed help right away for an illness or injury, and (4) saw a doctor within 15 minutes of their scheduled appointment time. Although all of these questions have a conceptual relationship to getting care right away, there is no particular reason to think that how well an office handles phone requests for information would be strongly related to how long patients wait in the waiting room to see a doctor. These items are put into the same index because the investigators grouped them

O'Muircheartaigh (1991) on Reinterviews to Estimate Simple Response Variance

O'Muircheartaigh (1991) examined data from reinterviews of the Current Population Survey (CPS) designed to estimate simple response variance.

Study design: About a week after CPS interviews, a different interviewer reinterviewed a 1/18 sample of respondents. Any eligible respondent reported the data on each occasion. The gross difference rate (GDR) is merely the likelihood of a discrepancy between the first and second report.

Findings: There are higher GDRs when persons report for themselves than about others, for reports about younger persons, and for reports made by nonheads of households.

Limitations of the study: The finding of higher response variance for self-reporters is limited by the fact that there was no random assignment of reporter status. Reporters tend to be those at home more often and thus tend to be unemployed. The study did not attempt to measure response biases, just instability over time. The reinterview is sometimes conducted by more senior interviewers using a different mode than the that of the first interview.

Impact of the study: The study identified systematic influences on stability of answers over repeated trials. It demonstrated the value of reinterview data when studying correlates of response variability in an ongoing survey. The finding about higher instability of self-reporters was plausibly explained by their tendency to undergo more change in their employment. This suggests lower reliability of reports for rapidly changing attributes.

together, not because they are measures of the same underlying process or phenomenon.

When the items are correlated, they do not correlate with one another at a particularly high level. The value of alpha, the measure of internal reliability described above, is only 0.58. However, the composite measure as a whole, formed by summing the responses to the four questions, turns out to be a highly significant predictor of how people rate their overall health care (0.57) and it was found to be highly reliable in providing a measure of this construct for health plans ($r = 0.94$). Thus, despite the fact that the items in the composite are not all measuring the same underlying construct, they do provide a reliable measure, one that is an important predictor of respondents' overall ratings of their health care.

In summary, both reinterviews and use of multiple items to measure reliability or simple response variance require assumptions that can be challenging to uphold. Nonetheless, both techniques are common and useful for survey researchers.

8.10 SUMMARY

Evaluating survey questions has two components. The first is determining whether the right questions are being asked, whether respondents understand them as intended and can answer them without undue difficulty, and whether the questions can be administered easily under field conditions. We described five methods for addressing these issues: expert reviews, focus groups, cognitive testing, field pretests, and split-ballot experiments. Almost all surveys use one or more of these methods in developing questionnaires. The different methods yield somewhat different information. The methods chosen for any particular survey are likely to depend on the specific issues of concern to the survey designers, the budget for the survey, and whether most or all of the questions have been used before. Unfortunately, we have little data on whether the assessments emerging from these techniques really pinpoint the most serious problems affecting the survey responses.

Statistical evaluations of survey questions measure the validity or reliability of the answers and the response variance and bias of summary statistics. There are two principal methods to estimate validity: comparing the survey responses to external data (such as records) or determining whether survey-derived measures follow theoretical expectations.

Reliability and simple response variance are commonly assessed by administering the same question to a respondent twice, once in the main interview and a second time during a reinterview. Another approach is to administer multiple items assessing the same construct in the same interview and to examine the consistency of the answers across the items.

Measuring response bias compares survey responses to some external data, either data for individual respondents or statistics on the target population.

Comparing survey responses to accurate external data generally provides the most useful estimates of the error in the survey data, but such external data are seldom available. (When they are available, it is often for special populations, such as the members of a particular HMO or persons who were hospitalized at a

specific hospital, who may not be fully representative of the populations in which the questions will be used.)

The more indirect methods for assessing validity generally provide less-convincing evidence regarding the level of measurement error. Although it is reassuring to be able to say that the answers to the questions have the predicted relationships to other variables, that does not give a quantified estimate of the level of measurement error. Still, such evidence is often the best we can do to evaluate the quality of the answers to survey questions.

KEYWORDS

behavior coding
cognitive interviewing
cognitive standards
content standards
Cronbach's alpha
expert review
focus group
gross difference rate
index of inconsistency
interviewer debriefing
pretest
protocol analysis
randomized experiment
record check study
reinterview
reliability
response bias
simple response variance
split-ballot experiment
usability standards
validity

FOR MORE IN-DEPTH READING

Alwin, D. (2007), *Margins of Error*, New York: Wiley.

Harkness, J., Vijver, F., and Mohler, P. (2002), *Cross-Cultural Survey Methods*, New York: Wiley.

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

Saris, W. and Gallhofer, I. (2007), *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, New York: Wiley.

Willis, G. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design,* Thousand Oaks, CA: Sage.

EXERCISES

1) Compare and contrast the benefits of three methods used in questionnaire development: cognitive testing, focus groups, and coding interviewer and respondent behavior during field pretests. By "benefits" is meant the acquisition of information about weaknesses in the design of the questionnaire that can be repaired prior to the main survey data collection. Name at least three ways the methods differ in their benefits.

2) Cognitively test the following questions with a friend or two, and think about how well you find they stand up to cognitive standards:

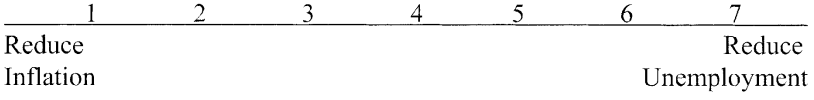
   a) What was your income in the past year?
   b) How often do you exercise—almost every day, several times a week, once a week, once a month, or less often?
   c) Do you favor or oppose universal health insurance in the United States?
   d) In the past year, about how many times did you use an ATM machine for any transaction?
   e) Consider the statement: "I am happier than usual these days." Would you say you strongly agree, generally agree, neither agree nor disagree, somewhat disagree, or strongly disagree with this statement?

3) Name two different ways to find out if a question designed to measure objective facts is producing valid data

4) Describe two different kinds of analyses one could do to assess the validity of answers to questions designed to measure a subjective state, such as happiness or anxiety.

5) For each of the items below, design two probes to be used in a cognitive interview that might help to identify any possible problems with the questions.

   a) During the past four weeks, beginning [date four weeks ago] and ending today, have you done any exercise, including sports, physically active hobbies, and aerobic exercises, but not including any activities carried out as part of your job or in the course of ordinary housework?
   b) How many times each week do you have milk, butter, or other dairy products?
   c) Living where you do now and meeting the expenses you consider necessary, what would be the smallest income (before any deductions) you and your family would need to make ends meet each month?
   d) Some people feel the federal government should take action to reduce the inflation rate even if it means that unemployment would go up a lot. Others feel the government should take action to reduce the rate of unemployment even if it means the inflation rate would go up a lot. Where would you place yourself on this [seven point] scale?



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

6) For each of the following scenarios, list one pretesting technique that would most directly address the problem at hand, and state why that technique would be useful.

   a) You are beginning to draft questions for a new survey, and you need to know how your target population thinks and talks about the survey topic—what words they use, how they define those terms, and so on. What technique should you employ?
   b) Your primary concern for a survey about to go into the field is that the interviewer–respondent interaction be as standardized as possible. You found during your own very informal testing (administering the questionnaire to a handful of coworkers) that the interaction was somewhat awkward—you were sometimes interrupted with an answer before you had read all of the response categories to the respondent, other times you were asked to repeat questions, and in some cases you were asked what certain words meant (although no standard definition was available). What would be the best way to address this concern in regard to standardization of interviewer–respondent interaction during a pretest?
   c) You are very concerned about potential comprehension and recall problems for a questionnaire you have been asked to finalize for data collection. You will ultimately conduct a large-scale "dress rehearsal" field pretest several months from now, but something must be done before then to improve the questionnaire. Some questions have ambiguous wording, and could easily be interpreted in different ways by different people; other questions ask for information that seems rather difficult to recall for most people. In addition, you suspect that there may be different problems for various subgroups of the population (e.g., those with low levels of education, those from different ethnic groups, etc.). What should you do?

7) You have very limited funds for pretesting, and your client has given you a questionnaire that he/she thinks is "ready to go." Upon looking at it, you see major issues with question wording, flow of topics, and so on. You only have enough money to conduct one field pretest, and you need to convince your client that the questionnaire needs revision prior to the field pretest. You have very limited time in which to address the problems with the questionnaire before starting the field pretest. What should you do?

8) You are developing a questionnaire that asks about complicated topics over long periods of time, and you suspect there may be different problems for various subgroups of the population (e.g., those with low levels of education, those from different ethnic groups, etc.). You've been asked to finalize the questionnaire for data collection. You will ultimately conduct a large-scale "dress rehearsal" field pretest several months from now, but something must be done before then to improve the questionnaire. What should you do?

9) You have two subject matter specialists in your organization who have each written what they consider to be the "best" question on a particular topic.

Only one of these can be included in the final version of the survey. Both people are adamant that their version is better, and you are about to conduct a fairly large-scale field pretest before starting the main survey. How could you address this situation in a way that would appease both of your specialists?

10) Suppose you have used a multi-item series of questions to measure how happy respondents have been in the past week. Each of the four items asks respondents to describe themselves using words with similar meanings, such as joyful, cheerful, and upbeat. The average correlation among the answers to the four questions is 0.60.

    a) What is the value of Cronbach's alpha for this four-item index?
    b) What does Chronbach's alpha measure?

11) Describe briefly two techniques used to measure response bias. What are the pros and cons of each?

12) You conduct a reinterview study in which a new interviewer asks respondents a sample of the questions from a survey two weeks after the original survey. For some questions, the results are almost identical; for others there are some notable differences between the answers that respondents gave to the same questions two weeks apart.

    a) What are the technical terms used for the degree of correspondence between the two answers?
    b) What are four reasons that answers might differ?
    c) What are the implications for the validity of the answers as measures when the answers are not highly consistent?
    d) What are the implications for the validity of the answers as measures when the answers are highly consistent?