METHODS FOR TESTING AND EVALUATING SURVEY QUESTIONS

STANLEY PRESSER
University of Maryland
MICK P. COUPER
University of Michigan
JUDITH T. LESSLER
Research Triangle Institute
ELIZABETH MARTIN
U.S. Census Bureau
JEAN MARTIN
Office for National Statistics
JENNIFER M. ROTHGEB
U.S. Census Bureau
ELEANOR SINGER
University of Michigan

An examination of survey pretesting reveals a paradox. On the one hand, pretesting is the only way to evaluate in advance whether a questionnaire causes problems for interviewers or respondents. Consequently, both elementary textbooks and experienced researchers declare pretesting indispensable. On the other hand, most textbooks offer minimal, if any, guidance about pretesting methods, and published survey reports usually provide no information about whether questionnaires were pretested and, if so, how, and with what results. Moreover, until recently there was relatively little methodological research on pretesting. Thus pretesting's universally acknowledged importance has been honored more in the breach than in the practice, and not a great deal is known about many aspects of pretesting, including the extent to which pretests serve their intended purpose and lead to improved questionnaires.

Pretesting dates to the founding of the modern sample survey in the mid-1930s or shortly thereafter. The earliest references in scholarly journals are from 1940, by which time pretests apparently were well established. In that

This is a revised version of chapter 1 from Presser et al., 2004.

Public Opinion Quarterly, Vol. 68 No. 1 Pp. 109–130,
DOI:10.1093/poq/nfh008

year Katz reported, "The American Institute of Public Opinion [i.e., Gallup] and *Fortune* [i.e., Roper] pretest their questions to avoid phrasings which will be unintelligible to the public and to avoid issues unknown to the man on the street" (Katz 1940, p. 279).

Although the absence of documentation means we cannot be sure, our impression is that for much of survey research's history, there has been one conventional form of pretest. Conventional pretesting is essentially a dress rehearsal, in which interviewers receive training like that for the main survey and administer the questionnaire as they would during the survey proper. After each interviewer completes a handful of interviews, response distributions may be tallied, and there is a debriefing in which the interviewers relate their experiences with the questionnaire and offer their views about the questionnaire's problems.

Survey researchers have shown remarkable confidence in this approach. According to one leading expert, "It usually takes no more than 12–25 cases to reveal the major difficulties and weaknesses in a pretest questionnaire" (Sheatsley 1983, p. 226). This judgment is similar to that of another prominent methodologist, who maintained that "20–50 cases is usually sufficient to discover the major flaws in a questionnaire" (Sudman 1983, p. 181).

This faith in conventional pretesting is probably based on the common experience that a small number of conventional interviews often reveal numerous problems, such as questions that contain unwarranted suppositions, awkward wordings, or missing response categories. However, there is no scientific evidence to justify the confidence that this kind of pretesting identifies the major problems in a questionnaire.

Conventional pretests are based on the assumption that questionnaire problems will be signaled either by the answers that the questions elicit (e.g., "don't knows" or refusals), which will show up in response tallies, or by some other visible consequence of asking the questions (e.g., hesitation or discomfort in responding), which interviewers can describe during debriefing. However, as Cannell and Kahn (1953, p. 353) noted, "There are no exact tests for these characteristics." They go on to say, "The help of experienced interviewers is most useful at this point in obtaining subjective evaluations of the questionnaire." Similarly, Moser and Kalton (1971, p. 50) judged, "Almost the most useful evidence of all on the adequacy of a questionnaire is the individual fieldworker's [i.e., interviewer's] report on how the interviews went, what difficulties were encountered, what alterations should be made, and so forth." This emphasis on interviewer perceptions is nicely illustrated in Sudman and Bradburn's (1982, p. 49) advice for detecting unexpected word meanings: "A careful pilot test conducted by *sensitive* interviewers is the most direct way of discovering these problem words" (emphasis added).

Yet even if interviewers were extensively trained in recognizing problems with questions (as compared with receiving no special training at all, which is typical), conventional pretesting would still be ill suited to uncovering many

questionnaire problems. Certain kinds of problems will not be apparent from observing respondent behavior, and the respondents themselves may be unaware of the problems. For instance, respondents can misunderstand a closed question's intent without providing any indication of having done so. Moreover, because conventional pretests are almost always "undeclared" to the respondent, as opposed to "participating" (in which respondents are informed of the pretest's purpose; see Converse and Presser 1986), respondents are usually not asked directly about their interpretations or other problems the questions may have caused. As a result, undeclared conventional pretesting seems better designed to identify problems the questionnaire poses for interviewers, who know the purpose of the testing, than for respondents, who do not.

Furthermore, when conventional pretest interviewers do describe respondent problems, there are no rules for assessing their descriptions or for determining which problems that are identified ought to be addressed. Researchers typically rely on intuition and experience in judging the seriousness of problems and deciding how to revise questions that are thought to have flaws.

In recent decades a growing awareness of conventional pretesting's drawbacks has led to two interrelated changes. First, there has been a subtle shift in the goals of testing, from an exclusive focus on identifying and fixing overt problems experienced by interviewers and respondents to a broader concern for improving data quality so that measurements meet a survey's objectives. Second, new testing methods have been developed or adapted from other uses. These methods include cognitive interviews, behavior coding, response latency, vignette analysis, formal respondent debriefings, experiments, and statistical modeling.[1] The development of these methods raises issues of how they might best be used in combination, as well as whether they in fact lead to improvements in survey measurement. In addition, the adoption of computerized modes of administration poses special challenges for pretesting, as do surveys of special populations, such as children, establishments, and those requiring questionnaires in more than one language—all of which have greatly increased in recent years. We review these developments, drawing on the latest research presented in the first volume devoted exclusively to testing and evaluating questionnaires (Presser et al. 2004).

Cognitive Interviews

Ordinary interviews focus on producing codable responses to the questions. Cognitive interviews, by contrast, focus on providing a view of the processes

1. All the methods discussed in this article involve data collection to test a questionnaire. We do not treat focus groups (Bischoping and Dykema 1999) or ethnographic interviews (Gerber 1999), which are most commonly used at an early stage, before there is an instrument to be tested. Nor do we review evaluations by experts (Presser and Blair 1994), artificial intelligence (Graesser et al. 2000), or coders applying formal appraisal systems (Lessler and Forsyth 1996), none of which involve data collection from respondents.

elicited by the questions. Concurrent or retrospective "think-alouds" and/or probes are used to produce reports of the thoughts respondents have either as they answer the survey questions or immediately after. The objective is to reveal the thought processes involved in interpreting a question and arriving at an answer. These thoughts are then analyzed to diagnose problems with the question.

Although he is not commonly associated with cognitive interviewing, William Belson (1981) pioneered a version of this approach. In the mid-1960s Belson designed "intensive" interviews to explore seven questions respondents had been asked the preceding day during a regular interview administered by a separate interviewer. Respondents were first reminded of the exact question and the answer they had given to it. The interviewer then inquired, "When you were asked that question yesterday, exactly what did you think the question meant?" After nondirectively probing to clarify what the question meant to the respondent, interviewers asked, "Now tell me exactly how you worked out your answer from that question. Think it out for me just as you did yesterday… only this time say it aloud for me." Then, after nondirectively probing to illuminate how the answer was worked out, interviewers posed scripted probes about various aspects of the question. These probes differed across the seven questions and were devised to test hypotheses about problems particular to each of the questions. Finally, after listening to the focal question once more, respondents were requested to say how they would now answer it. If their answer differed from the one they had given the preceding day, they were asked to explain why (Appendix, pp. 194–97). Six interviewers, who received two weeks of training, conducted 265 audiotaped, intensive interviews with a cross-section sample of London, England residents. Four analysts listened to the tapes and coded the incidence of various problems.

These intensive interviews differed in a critical way from today's cognitive interviews, which integrate the original and follow-up interviews in a single administration with one interviewer. Belson assumed that respondents could accurately reconstruct their thoughts from an interview conducted the previous day, which is inconsistent with what we now know about the validity of self-reported cognitive processes. However, in many respects, Belson moved considerably beyond earlier work, such as Cantril and Fried (1944), which used just one or two scripted probes to assess respondent interpretations of survey questions. Thus, it is ironic that Belson's approach had little impact on pretesting practices, an outcome possibly due to its being so labor-intensive.

The pivotal development leading to a role for cognitive interviews in pretesting did not come until two decades later with the Cognitive Aspects of Survey Methodology (CASM) conference (Jabine et al. 1984). Particularly influential was Loftus's (1984) postconference analysis of how respondents answered survey questions about past events, in which she drew on the think-aloud technique used by Herbert Simon and his colleagues to study problem solving (Ericsson and Simon 1980). Subsequently, a grant from Murray Aborn's program at the National Science Foundation to Monroe Sirken

supported both research on the technique's utility for understanding responses to survey questions (Lessler, Tourangeau, and Salter 1989) and the creation at the National Center for Health Statistics (NCHS) in 1985 of the first "cognitive laboratory," where the technique could routinely be used to pretest questionnaires (e.g., Royston and Bercini 1987).

Similar cognitive laboratories were soon established by other U.S. statistical agencies and survey organizations.[2] The labs' principal, but not exclusive, activity involved cognitive interviewing to pretest questionnaires. Facilitated by special exemptions from Office of Management and Budget survey clearance requirements, pretesting for U.S. government surveys increased dramatically through the 1990s (Martin, Schechter, and Tucker 1999). At the same time, the labs took tentative steps toward standardizing and codifying their practices in training manuals (e.g., Willis 1994) or protocols for pretesting (e.g., DeMaio et al. 1993).

Although there is now general agreement about the value of cognitive interviewing, no consensus has emerged about best practices, such as whether (or when) to use think-alouds versus probes, whether to employ concurrent or retrospective reporting, and how to analyze and evaluate results. In part this is due to the paucity of methodological research examining these issues, but it is also due to a lack of attention to the theoretical foundation for applying cognitive interviews to survey pretesting.

As Willis (2004) notes, Ericsson and Simon (1980) argued that verbal reports are more likely to be veridical if they involve information a person has available in short-term (as opposed to long-term) memory, and if the verbalization itself does not fundamentally alter thought processes (e.g., does not involve further explanation). Thus some survey tasks (for instance, nontrivial forms of information retrieval) may be well suited to elucidation in a think-aloud interview. However, the *general* use of verbal report methods to target cognitive processes involved in answering survey questions is difficult to justify, especially for tasks (such as term comprehension) that do not satisfy the conditions for valid verbal reports. Willis also notes that the social interaction involved in interviewer-administered cognitive interviews may violate a key assumption posited by Ericsson and Simon for use of the method.

Research has demonstrated various problems with the methods typically used to conduct cognitive interview pretests. Beatty (2004), for example, found that certain kinds of probes produce difficulties that respondents would not otherwise experience. His analysis of a set of cognitive interviews indicated that respondents who received re-orienting probes (asking for an answer) had little difficulty choosing an answer, whereas those who received elaborating

2. Laboratory research to evaluate self-administered questionnaires was already underway at the Census Bureau before the 1980 census (Rothwell 1983, 1985). Although inspired by marketing research rather than cognitive psychology, this work, in which observers encouraged respondents to talk aloud as they filled out questionnaires, foreshadowed cognitive interviewing. See also Hunt, Sparkman, and Wilcox 1982.

probes (asking for further information) had considerable difficulty. Beatty also found that, aside from reading the questions, cognitive probes (those traditionally associated with cognitive interviews, such as "What were you thinking?" "How did you come up with that?" or "What does [term] mean to you?") accounted for less than one-tenth of all interviewer utterances. Over nine-tenths consisted of confirmatory probes (repeating something the respondent said, in a request for confirmation), expansive probes (requests for elaboration, such as "Tell me more about that"), functional remarks (repetition or clarification of the question, including re-orienting probes), and feedback (e.g., "thanks; that's what I want to know" or "I know what you mean"). Thus cognitive interview results appear to be importantly shaped by the interviewers' contributions, which may not be well focused in ways that support the inquiry. As one way to deal with this problem, Beatty recommended that cognitive interviewers be trained to recognize distinctions among probes and the situations in which each ought to be employed.

Conrad and Blair (2004) argue that verbal report quality should be assessed in terms of problem detection and problem repair, which are the central goals of cognitive interviewing. They designed an experimental comparison of two different cognitive interviewing approaches: one, uncontrolled, using the unstandardized practices of four experienced cognitive interviewers; the other, more controlled, using four less experienced interviewers trained to probe only when there were explicit indications the respondent was experiencing a problem. The conventional cognitive interviews identified many more problems than did the conditional probe interviews.

As in Beatty (2004), however, more problems did not mean higher-quality results. Conrad and Blair assessed the reliability of problem identification in two ways: by inter-rater agreement among a set of trained coders who reviewed transcriptions of the taped interviews, and by agreement between coders and interviewers. Overall, agreement was quite low, consistent with the finding of some other researchers about the reliability of cognitive interview data (Presser and Blair 1994). But reliability was higher for the conditional probe interviews than for the conventional ones. (This may be partly due to the conditional probe interviewers having received training in what should be considered a "problem," compared to the conventional interviewers who were provided no definition of what constituted a "problem.") Furthermore, as expected, conditional interviewers probed much less often than conventional interviewers, but more of their probes were in cases associated with the identification of a problem. Thus we need to rethink what interviewers do in cognitive interviews.

The importance of this rethinking is underscored by DeMaio and Landreth (2004), who conducted an experiment in which three different organizations were commissioned to have two interviewers each conduct five cognitive interviews of the same questionnaire using whatever methods were typical for the organization, and then deliver a report identifying problems in the questionnaire as well as a revised questionnaire addressing the problems. In

addition, expert reviews of the original questionnaire were obtained from three individuals who were not involved in the cognitive interviews. Finally, another set of cognitive interviews was conducted by a fourth organization to test the revised questionnaires.

The three organizations reported considerable diversity on many aspects of the interviews, including location (respondent's home versus research lab), interviewer characteristics (field interviewer versus research staff), question strategy (think-aloud versus probes), and data source (review of audiotapes versus interviewer notes and recollections). This heterogeneity is consistent with the findings of Blair and Presser (1993), but it is even more striking given the many intervening years in which some uniformity of practice might have emerged. It does, however, mean that differences in the results across the organizations cannot be attributed to any one factor.

There was variation across the organizations in both the number of questions identified as having problems and the total number of problems identified. Moreover, there was only modest overlap across the organizations in the particular problems diagnosed. Likewise, the cognitive interviews and the expert reviews overlapped much more in identifying which questions had problems than in identifying what the problems were. The organization that identified the fewest problems also showed the lowest agreement with the expert panel. This organization was the only one that did not review the audiotapes in evaluating the results, which suggests that relying solely on interviewer notes and memory leads to error.[3] However, the findings from the tests of the revised questionnaires did not identify one organization as consistently better or worse than the others.

In sum, research on cognitive interviews has begun to reveal how the methods used to conduct the interviews shape the data produced. Yet much more work is needed to provide a foundation for optimal cognitive interviewing.

Supplements to Conventional Pretests

Unlike cognitive interviews, which are completely distinct from conventional pretests, other testing methods that have been developed may be implemented as add-ons to conventional pretests (or as additions to a survey proper). These include behavior coding, response latency, formal respondent debriefings, and vignettes.

Behavior coding was developed in the 1960s by Charles Cannell and his colleagues at the University of Michigan Survey Research Center, and it can be used to evaluate both interviewers and questions. Its early applications were almost entirely focused on interviewers, so it had no immediate impact on pretesting practices. In the late 1970s and early 1980s a few European researchers adopted behavior coding to study questions, but it was not applied

3. Bolton and Bronkhorst (1996) describe a computerized approach to evaluating cognitive interview results, which should reduce error even further.

to pretesting in the United States until the late 1980s (Oksenberg, Cannell, and Kalton's 1991 article describes behavior coding as one of two "new strategies for pretesting questions").

Behavior coding involves monitoring interviews or reviewing taped interviews (or transcripts) for a subset of the interviewer's and respondent's verbal behavior in the question asking and answering interaction. Questions marked by high frequencies of certain behaviors (e.g., the interviewer did not read the question verbatim or the respondent requested clarification) are seen as needing repair.

Van der Zouwen and Smit (2004) describe an extension of behavior coding that draws on the sequence of interviewer and respondent behaviors, not just the frequency of the individual behaviors. Based on the sequence of a question's behavior codes, an interaction is coded as either paradigmatic (the interviewer read the question correctly, the respondent chose one of the offered alternatives, and the interviewer coded the answer correctly), problematic (the sequence was nonparadigmatic, but the problem was solved; e.g., the respondent asked for clarification and then chose one of the offered alternatives), or inadequate (the sequence was nonparadigmatic, and the problem was not solved). Questions with a high proportion of nonparadigmatic sequences are identified as needing revision.

Van der Zouwen and Smit compared the findings from this approach in a survey of the elderly with the findings from basic behavior coding and from four "ex ante" methods—that is, methods not entailing data collection: a review by five methodology experts; reviews by the authors guided by two different questionnaire appraisal coding schemes; and the "quality predictor" developed by Saris and his colleagues, which we describe in the "statistical modeling" section below. The two methods based on behavior codes produced very similar results, as did three of the four ex ante methods—but the two sets of methods identified very different problems. As Van der Zouwen and Smit observe, the ex ante methods point out what *could* go wrong with the questionnaire, whereas the behavior codes and sequence analyses reveal what actually *did* go wrong.

Another testing method based on observing behavior involves the measurement of "response latency," the time it takes a respondent to answer a question. Since most questions are answered rapidly, latency measurement requires the kind of precision (to fractions of a second) that is almost impossible without computers. Thus it was not until after the widespread diffusion of computer-assisted survey administration in the 1990s that the measurement of response latency was introduced as a testing tool (Bassili and Scott 1996).

Draisma and Dijkstra (2004) used response latency to evaluate the accuracy of respondents' answers and, therefore, indirectly to evaluate the questions themselves. The authors reasoned that longer delays signal respondent uncertainty, and they tested this idea by comparing the latency of accurate and inaccurate answers (with accuracy determined by information from another source). In addition, they compared the performance of response latency to that of several other indicators of uncertainty.

In a multivariate analysis, both longer response latencies and the respondents' expressions of greater uncertainty about their answers were associated with inaccurate responses. Other research (Martin 2004; Schaeffer and Dykema 2004) reports no relationship (or even, in some studies, an inverse relationship) between respondents' confidence or certainty and the accuracy of their answers. Thus future work needs to develop a more precise specification of the conditions in which different measures of respondent uncertainty are useful in predicting response error.

Despite the fact that the interpretation of response latency is less straightforward than that of other measures of question problems (lengthy times may indicate careful processing, as opposed to difficulty), the method appears sufficiently promising to encourage its further use. This is especially so as the ease of collecting latency information means it could be routinely included in computer-assisted surveys at very low cost. The resulting collection of data across many different surveys would facilitate improved understanding of the meaning and consequences of response latency and of how it might best be combined with other testing methods, such as behavior coding, to enhance the diagnosis of questionnaire problems.

Unlike behavior coding and response latency, which are "undeclared" testing methods, respondent debriefings are a "participating" method, which informs the respondent about the purpose of the inquiry. Such debriefings have long been recommended as a supplement to conventional pretest interviews (Kornhauser 1951, p. 430), although they most commonly have been conducted as unstructured inquiries improvised by interviewers. Martin (2004) shows how implementing debriefings in a standardized manner can reveal both the meanings of questions and the reactions respondents have to the questions. In addition, she demonstrates how debriefings can be used to measure the extent to which questions lead to missed or misreported information.

Martin (2004) also discusses vignettes—hypothetical scenarios that respondents evaluate—which may be incorporated in either undeclared or participating pretests. Vignette analysis appears well suited to (1) explore how people think about concepts; (2) test whether respondents' interpretations of concepts are consistent with those that are intended; (3) analyze the dimensionality of concepts; and (4) diagnose other question wording problems. Martin offers evidence of vignette analysis's validity by drawing on evaluations of questionnaire changes made on the basis of the method.

The research we have reviewed suggests that the various supplements to conventional pretests differ in the kinds of problems they are suited to identify, their potential for diagnosing the nature of a problem and thereby for fashioning appropriate revisions, the reliability of their results, and the resources needed to conduct them. It appears, for instance, that formal respondent debriefings and vignette analysis are more apt than behavior coding and response latency to identify certain types of comprehension problems. Yet we do not have good estimates of many of the ways the methods differ. The

implication is not only that we need research explicitly designed to make such comparisons, but also that multiple testing methods are probably required in many cases to ensure that respondents understand the concepts underlying questions and are able and willing to answer them accurately (for good examples of multimethod applications, see Kaplowitz, Lupi, and Hoehn [2004] and Schaeffer and Dykema [2004]).

Experiments

Both supplemental methods to conventional pretests and cognitive interviews identify questionnaire problems and lead to revisions designed to address the problems. To determine whether the revisions are improvements, however, there is no substitute for experimental comparisons of the original and revised items. Such experiments are of two kinds. First, the original and revised items can be compared using the testing method(s) that identified the problem(s). Thus, if cognitive interviews showed respondents had difficulty with an item, the item and its revision can be tested in another round of cognitive interviews in order to confirm that the revision shows fewer such problems than the original. The interpretation of results from this kind of experiment is usually straightforward, though there is no assurance that observed differences will have any effect on survey estimates.

Second, original and revised items can be tested to examine what, if any, difference they make for a survey's estimates. The interpretation from this kind of experiment is sometimes less straightforward, but such split-sample experiments have a long history in pretesting. Indeed, they were the subject of one of the earliest articles devoted to pretesting (Sletto 1950), although the experiments it described dealt with the impact on cooperation to mail surveys of administrative matters such as questionnaire length, nature of the cover letter's appeal, use of follow-up postcards, and questionnaire layout. None of the examples concerned question wording.

Fowler (2004) describes three ways to evaluate the results of experiments that compare question wordings: differences in response distributions, validation against a standard, and usability, as measured, for instance, by behavior coding. He illustrates how cognitive interviews and experiments are complementary: the former identify potential problems and propose solutions, and the latter test the impact of the solutions. As he argues, experimental evidence is essential in estimating whether different question wordings affect survey results, and if so, by how much.

Fowler focuses on comparisons of single items that vary in only one way. Experiments can also be employed to test versions of entire questionnaires that vary in multiple, complex ways, as described by Moore et al. (2004). These researchers revised the Survey of Income and Program Participation (SIPP) questionnaire to meet three major objectives: to minimize response

burden and thereby decrease both unit and item nonresponse; to reduce "seam bias" reporting errors; and to introduce questions about new topics. Then, to assess the effects of the revisions before switching to the new questionnaire, an experiment was conducted in which respondents were randomly assigned to either the new or old version.

Both item nonresponse and seam bias were lower with the new questionnaire, and, with one exception, the overall estimates of income and assets (key measures in the survey) did not differ between versions. On the other hand, unit nonresponse reductions were not obtained (in fact, in initial waves, nonresponse was higher for the revised version), and the new questionnaire took longer to administer. Moore et al. note that these results may have been caused by two complicating features of the experimental design. First, experienced SIPP interviewers were used for both the old and new instruments. The interviewers' greater comfort level with the old questionnaire (some reported being able to "administer it in their sleep") may have contributed to their administering it more quickly than the new questionnaire and persuading more respondents to cooperate with it. Second, the addition of new content to the revised instrument may have more than offset the changes that were introduced to shorten the interview.

Tourangeau (2004) argues that the practical consideration that leads many experimental designs to compare packages of variables, as in the SIPP case, hampers the science of questionnaire design. Because the SIPP research experimented with a package of variables, it could estimate the overall effect of the redesign, which is vital to the SIPP sponsors, but not estimate the effects of individual changes, which is vital to an understanding of the effects of questionnaire features (and therefore to sponsors of other surveys making design changes). Relative to designs comparing packages of variables, factorial designs allow inference not only about the effects of particular variables, but about the effects of interactions between variables as well. Greater use of factorial designs (as well as more extensive use of laboratory experiments, for which Tourangeau also argues because they are usually much cheaper than field experiments) is therefore needed.

Statistical Modeling

Questionnaire design and statistical modeling are usually thought of as worlds apart. Researchers who specialize in questionnaires tend to have rudimentary statistical understanding, and those who specialize in statistical modeling generally have little appreciation for question wording. This is unfortunate, as the two should work in tandem for survey research to progress. Moreover, the "two worlds" problem is not inevitable. In the early days of survey research, Paul Lazarsfeld, Samuel Stouffer, and their colleagues made fundamental contributions to both questionnaire design and statistical analysis (e.g.,

Stouffer et al. 1950). Thus it is fitting that one recent development to evaluate questionnaires draws on a technique, "latent class analysis" (LCA), rooted in Lazarsfeld's work.

Paul Biemer (2004) shows how LCA may be used to estimate the error associated with questions when the questions have been asked of the same respondents two or more times. Yet, as Biemer notes, LCA depends heavily on an assumed model, and there is usually no direct way to evaluate the model assumptions. He recommends that rather than relying on a single statistical method for evaluating questions, multiple methods ought to be employed.

Whereas research like Biemer's focuses on individual survey questions, psychometricians have long focused on the properties of scales composed of many items. Traditionally, applications of classical test theory have provided little information about the performance of the separate questions. Reeve and Mâsse (2004) describe how item response theory (IRT) models can assess the degree to which different items discriminate among respondents who have the same value on a trait. The power of IRT to identify the discriminating properties of specific items allows researchers to design shorter scales that do a better job of measuring constructs. Even greater efficiency can be achieved by using IRT methods to develop computer adaptive tests (CAT). With CAT, a respondent is presented a question near the middle of the scale range, and an estimate of his total score is constructed based on his response. Another item is then selected based on that estimate, and the process is repeated. At each step, the precision of the estimated total score is computed, and when the desired precision is reached, no more items are presented.

Both latent class analysis and item response theory models require large numbers of cases and thus are relatively expensive to conduct. By contrast no new data collection is required to make use of a statistical modeling approach first proposed by Frank Andrews. Andrews (1984) applied the multitrait, multimethod (MTMM) measurement strategy (Campbell and Fiske 1959) to estimate the reliability and validity of a sample of questionnaire items, and he suggested the results could be used to characterize the reliability and validity of question types. Following his suggestion, Saris, Van der Veld, and Gallhofer (2004) created a data base of MTMM studies that provides estimates of reliability and validity for 1,067 questionnaire items. They then developed a coding system to characterize the items according to the nature of their content, complexity, type of response scale, position in the questionnaire, data collection mode, sample type, and the like. Two large regression models in which these characteristics were the independent variables and the MTMM reliability or validity estimates were the dependent variables provide estimates of the effect on the reliability or validity of the question characteristics. New items can be coded (aided by the authors' software) and the prediction equation (also automated) used to estimate their quality. Although more MTMM data are needed to improve the models, and—even more importantly—the

model predictions need to be tested in validation studies, such additional work promises a significant payoff for evaluating questions.

Mode of Administration

The introduction of computer technology has changed many aspects of questionnaires. On the one hand, the variety of new modes—beginning with computer-assisted telephone interviewing (CATI), but soon expanding to computer-assisted personal interviewing (CAPI) and computer-assisted self-interviewing (CASI)—has expanded our ability to measure a range of phenomena more efficiently and with improved data quality (Couper et al. 1998). On the other hand, the continuing technical innovations—including audio-CASI, interactive voice response, and the Internet—present many challenges for questionnaire design.

The proliferation of data collection modes has at least three implications for the evaluation and testing of survey instruments. One implication is the mounting recognition that answers to survey questions may be affected by the mode in which the questions are asked. Thus, testing methods must take into consideration the delivery mode. A related implication is that survey instruments consist of much more than words, e.g., their layout and design, logical structure and architecture, and the technical aspects of the hardware and software used to deliver them. All of these elements need to be tested, and their possible effects on measurement error explored. A third implication is that survey instruments are ever more complex and demand ever-expanding resources for testing. The older methods that relied on visual inspection to test flow and routing are no longer sufficient. Newer methods must be found to facilitate the testing of instrument logic, quite aside from the wording of individual questions. In sum, the task of testing questionnaires has greatly expanded.

With the growing complexity of computer-assisted survey instruments and the expanding range of design features available, checking for programming errors has become an increasingly costly and time-consuming part of the testing process, often with no guarantee of complete success. Much of this testing can be done effectively and efficiently only by machine, but existing software is often not up to the task (Cork et al. 2003; Tarnai and Moore 2004).

The visual presentation of information to the interviewer, as well as the design of auxiliary functions used by the interviewer in computer-assisted interviewing, are critical to creating effective instruments. Thus testing for usability can be as important as testing for programming errors. As Hansen and Couper (2004) argue, computerized questionnaires require interviewers to manage two interactions, one with the computer and another with the respondent, and the goal of good design must therefore be to help interviewers manage both interactions to optimize data quality. Hansen and Couper provide illustrations of the ways in which usability testing assists in achieving this end.

A focus on question wording is insufficient even in the technologically simple paper-and-pencil mode. Dillman and Redline (2004) demonstrate how cognitive interviews may be adapted to explore the various aspects of visual language in self-administered questionnaires. They also show how the results of cognitive interviews can aid in the interpretation of split-sample field experiments.

Web surveys require testing of aspects unique to that mode, such as respondents' monitor display properties, the presence of browser plug-ins, and features of the hosting platform that define the survey organization's server. In addition to testing methods used in other modes, Baker, Crawford, and Swinehart (2004) recommend evaluations based on process data that are easily collected during Web administration (e.g., response latencies, backups, entry errors, and breakoffs). Like Tarnai and Moore (2004), Baker, Crawford, and Swinehart underscore the importance of automated testing tools, and, consistent with Dillman and Redline (2004) and Hansen and Couper (2004), they emphasize that the testing of Web questionnaires must focus on their visual aspects.

Special Populations

Surveys of children, establishments, and populations that require questionnaires in multiple languages pose special design problems. Thus, pretesting is still more vital in these cases than it is for surveys of adults interviewed with questionnaires in a single language. Remarkably, however, pretesting has been even further neglected for such surveys than for "ordinary" ones. As a result, the methodological literature on pretesting is even sparser for these cases than for monolingual surveys of adults.

Willimack et al. (2004) describe distinctive characteristics of establishment surveys that have made questionnaire pretesting uncommon. Establishment surveys tend to be mandatory, to rely on records, and to target populations of a few very large organizations, which are included with certainty, and many smaller ones, which are surveyed less often. These features seem to have militated against adding to the already high respondent burden by conducting pretests. In addition, because establishment surveys are disproportionately designed to measure change over time, questionnaire changes are rare. Finally, establishment surveys tend to rely on post-collection editing to correct data.

Willimack et al. outline various ways to improve the design and testing of establishment questionnaires. In addition to greater use of conventional methods, they recommend strategies like focus groups, site visits, record-keeping studies, and consultation with subject area specialists and other stakeholders. They also suggest making better use of ongoing quality evaluations and reinterviews, as well as more routine documentation of respondents' feedback, to provide diagnoses of questionnaire problems. Finally, they recommend that

tests be embedded in existing surveys so that proposed improvements can be evaluated without increasing the burden.

In "Pretesting Questionnaires for Children and Adolescents," De Leeuw, Borgers, and Smits (2004) review studies of children's cognitive development for guidance about the kinds of questions and cognitive tasks that can be asked of children of different ages. The evidence suggests that 7 years old is about the earliest age at which children can be interviewed with structured questionnaires, although the ability to handle certain kinds of questions (e.g., hypothetical ones) is acquired only later. The authors discuss how various pretesting methods, including focus groups, cognitive interviews, observation, and debriefing, can be adapted to accommodate children of different ages, and they provide examples of pretests that used these methods with children.

Questionnaire translation has always been basic to cross-national surveys, and recently it has become increasingly important for national surveys as well. Some countries (e.g., Canada, Switzerland, and Belgium) must administer surveys in multiple languages by law. Other nations are translating questionnaires as a result of growing numbers of immigrants. In the United States, for instance, the population 18 years and older that speaks a language at home other than English increased from 13.8 percent in 1990 to 17.8 percent in 2000. Moreover, by 2000, 4.4 percent of U.S. adults lived in "linguistically isolated" households, those in which all the adults spoke a language other than English, and none spoke English "very well" (U.S. Census Bureau 2003).

Despite its importance, Smith (2004) reports that "no aspect of cross-national survey research has been less subjected to systematic, empirical investigation than translation." He describes sources of non-equivalence in translated questions and discusses the problems involved in translating response scales or categories so they are equivalent. He then outlines several strategies to address problems arising from noncomparability across languages: asking multiple questions about a concept (e.g., well-being) with different terms in each question (e.g., satisfaction versus happiness), so that translation problems with a single term do not result in measurement error for all the items; using questions that are equivalent across cultures and languages as well as those that are culture-specific; and conducting special studies to calibrate scale terms.

Harkness, Pennell, and Schoua-Glusberg (2004) offer guidance on procedures and protocols for translation and assessment. They envision a more rigorous process of "translatology" than the ad hoc practices common to most projects. They emphasize the need for appraisals of the translated text (and hence do not believe back-translation is adequate), and they argue that the quality of translations, as well as the performance of the translated questions as survey questions, must be assessed. Finally, they recommend team approaches that bring different types of expertise to bear on the translation, and they suggest ways to organize the effort of translation, assessment, and documentation (the last of which is particularly important for interpreting results after a survey is completed).

Effects of Testing

Does pretesting lead to better measurement? We know of only one study that unambiguously addresses this question. Forsyth, Rothgeb, and Willis (2004) assessed whether pretesting (a) predicts data collection problems and (b) improves survey outcomes. The authors used three methods—informal expert review, appraisal coding, and cognitive interviews—to identify potential problems in a pretest of a questionnaire consisting of 83 items. The 12 questions diagnosed most consistently by the three methods as having problems were then revised to address the problems. Finally, a split-sample field experiment was conducted to compare the original and revised items. The split-sample interviews were behavior coded, and the interviewers were asked to evaluate the questionnaires after completing the interviews.

The versions of the original questions identified in the pretest as particularly likely to pose problems for interviewers were more likely to show behavior-coded interviewing problems in the field and to be identified by interviewers as having posed problems for them. Similarly, the questions identified by the pretest as posing problems for respondents resulted in more respondent problems, according to both the behavior coding and the interviewer ratings. Item nonresponse was also higher for questions identified by the pretest as presenting either recall or sensitivity problems than for questions not identified as having those problems. Thus the combination of pretesting methods was a good predictor of the problems the items would produce in the field.

However, the revised questions generally did not appear to outperform the original versions. The item revisions had no effect on the frequency of behavior-coded interviewer and respondent problems. And while interviewers did rate the revisions as posing fewer respondent problems, they rated them as posing more interviewer problems. The authors suggest various possible explanations for this outcome, including their selection of only questions diagnosed as most clearly problematic, which often involved multiple problems that required complex revisions to address. In addition, the revised questions were not subjected to another round of testing using the three methods that originally identified the problems to confirm that the revisions were appropriate. Nonetheless, the results are chastening, as they suggest that we have much better tools for diagnosing questionnaire problems than for fixing them.

An Agenda for the Future

Different pretesting methods, and different ways of carrying out the same method, influence the numbers and types of problems identified. Consistency among methods is often low, and the reasons for this need more investigation. One possibility is that, in their present form, some of the methods are unreliable. But two other possibilities are also worth exploring. First, lack of consistency

may occur because the methods are suited for identifying different problem types. For example, comprehension problems that occur with no disruption in the question asking and answering process are unlikely to be picked up by behavior coding. Thus, we should probably expect only partial overlap in the problems identified by different pretesting methods. Second, inconsistencies may reflect a lack of consensus among researchers, cognitive interviewers, or coders about what is regarded as a problem. For example, is it a problem if a question is awkward to ask but obtains accurate responses, or is it only a problem if the question obtains erroneous answers? The kinds and severity of problems that a questionnaire pretest (or methodological evaluation) aims to identify are not always clear, and this lack of specification may contribute to the inconsistencies that have been observed.

In exploring such inconsistencies, the cross-organization approach used by DeMaio and Landreth (2004; see also Martin, Schechter, and Tucker 1999) holds promise not only of leading to greater standardization, and therefore to higher reliability, but to enhancing our understanding of which methods are appropriate in different circumstances and for different purposes.

It is also clear that problem *identification* does not necessarily point to problem *solution* in any obvious or direct way. For instance, Forsyth, Rothgeb, and Willis (2004) and Schaeffer and Dykema (2004) used pretesting to identify problems that were then addressed by revisions, only to find in subsequent field studies that the revisions either did not result in improvements or created new problems. The fact that we are better able to identify problems than to formulate solutions underscores the desirability of additional testing after questionnaires have been revised.

Four general recommendations seem particularly important to us for advancing questionnaire testing and evaluation. These involve

1. the connection between problem identification and measurement error;
2. the impact of testing methods on survey costs;
3. the role of basic research and theory in guiding the repair of question flaws; and
4. the development of a data base to facilitate cumulative knowledge.

First, we need studies that examine the connection between problem diagnosis and measurement error. A major objective of testing is to reduce measurement error, yet we know little about the degree to which error is predicted by the various problem indicators at the heart of the different testing methods. Draisma and Dijkstra (2004) and Schaeffer and Dykema (2004) are unusual in making use of external validation in this way. Other research has taken an indirect approach, by examining the link between problem diagnosis and specific response patterns (for example, missing data, or "seam bias"), on the assumption that higher or lower levels are more accurate. But inferences based on indirect approaches must be more tentative than those based on

direct validation (e.g., record-check studies). With appropriately designed validation studies, we might be better able to choose among techniques for implementing particular methods, evaluate the usefulness of different methods for diagnosing different kinds of problems, and understand how much pretesting is "enough." We acknowledge, however, that validation data are rarely available and are themselves subject to error. Thus another challenge for future research is to develop further indicators of measurement error that can be used to assess testing methods.

Second, we need information about the impact of different testing methods on survey costs. The cost of testing may be somewhat offset, completely offset, or even more than offset (and therefore reduce the total survey budget), depending on whether the testing results lead to the identification (and correction) of problems that affect those survey features—e.g., interview length, interviewer training, and post-survey data processing—that have implications for cost. Although we know something about the direct costs of various testing methods, we know almost nothing about how the methods differ in their impact on overall costs. Thus a key issue for future research is to estimate how different testing methods perform in identifying the kinds of problems that increase survey costs.

Third, since improved methods for diagnosing problems are mainly useful to the extent that we can repair the problems, we need more guidance in making repairs. As a result, advances in pretesting depend partly on advances in the science of asking questions (Schaeffer and Presser 2003). Such a science involves basic research into the question and answer process that is theoretically motivated (Krosnick and Fabrigar forthcoming; Sudman, Bradburn, and Schwarz 1996; Tourangeau, Rips, and Rasinski 2000). But this is a two-way street. On the one hand, pretesting should be guided by theoretically motivated research into the question and answer process. On the other hand, basic research and theories of the question and answer process should be shaped by both the results of pretesting and developments in the testing methods themselves, e.g., the question taxonomies, or classification typologies, used in questionnaire appraisal systems (Lessler and Forsyth 1996), and the kind of statistical modeling described by Saris, Van der Veld, and Gallhofer (2004). In particular, pretesting's focus on aspects of the response tasks that can make it difficult for respondents to answer accurately ought to inform theories of the connection between response error and the question and answer process.

Finally, we need improved ways to accumulate knowledge across pretests. This will require greater attention to documenting what is learned from pretests of individual questionnaires. One of the working groups at the Second Advanced Seminar on the Cognitive Aspects of Survey Methodology (Sirken et al. 1999, p. 56) suggested that survey organizations archive, in a central repository, the cognitive interviews they conduct, including the items tested, the methods used, and the findings produced. As that group suggested, this would "facilitate systematic research into issues such as: What characteristics of questions are identified by cognitive interviewing as engendering particular

problems? What testing features are associated with discovering different problem types? What sorts of solutions are adopted in response to various classes of problems?" We believe this recommendation should apply to *all* methods of pretesting. Establishing a pretesting archive on the Web would not only facilitate research on questionnaire evaluation; it would also serve as an invaluable resource for researchers developing questionnaires for new surveys.[4]

References

Andrews, Frank. 1984. "Construct Validity and Error Components of Survey Measures." *Public Opinion Quarterly* 48:409–42.

Baker, Reginald, Scott Crawford, and Janice Swineheart. 2004. "Development and Testing of Web Questionnaires." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Bassili, John and Stacey Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60:390–99.

Beatty, Paul. 2004. "The Dynamics of Cognitive Interviewing." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Belson, William. 1981. *The Design and Understanding of Survey Questions*. London: Gower.

Biemer, Paul. 2004. "Modeling Measurement Error to Identify Flawed Questions." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Bischoping, Katherine, and Jennifer Dykema. 1999. "Towards a Social Psychological Program for Improving Focus Group Methods of Developing Questionnaires." *Journal of Official Statistics* 15:495–516.

Blair, Johny, and Stanley Presser. 1993. "Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice." *Proceedings of the Section on Survey Research Methods of the American Statistical Association* 370–375.

Bolton, Ruth, and Tina Bronkhorst. 1996. "Questionnaire Pretesting: Computer-Assisted Coding of Concurrent Protocols." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz and Seymour Sudman, pp. 37–64. San Francisco: Jossey-Bass.

Campbell, Donald, and Donald Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait Multimethod Matrices." *Psychological Bulletin* 56:81–105.

Cannell, Charles, and Robert Kahn. 1953. "The Collection of Data by Interviewing." In *Research Methods in the Behavioral Sciences*, ed. Leon Festinger and Daniel Katz, pp. 327–80. New York: Dryden.

Cantril, Hadley, and Edrita Fried. 1944. "The Meaning of Questions." In *Gauging Public Opinion*, ed. Hadley Cantril, pp. 3–22. Princeton, NJ: Princeton University Press.

Conrad, Fred, and Johny Blair. 2004. "Data Quality in Cognitive Interviews." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Converse, Jean, and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Newbury Park, CA: Sage.

4. Many Census Bureau pretest reports are available online at www.census.gov/srd/www/byyear.html, and many other pretest reports may be found in the Proceedings of the American Statistical Association Survey Research Methods Section and the American Association for Public Opinion Research available at www.amstat.org/sections/srms/proceedings. But neither site is easily searchable, and the reports often contain incomplete information about the procedures used.

Cork, Daniel, Michael Cohen, Robert Groves, and William Kalsbeek, eds. 2003. *Survey Automation: Report and Workshop Proceedings*. Washington, DC: National Academies Press.

Couper, Mick, Reginald Baker, Jelke Bethlehem, Cynthia Clark, Jean Martin, William Nicholls II, and James O'Reilly. 1998. *Computer-Assisted Survey Information Collection*. New York: Wiley.

De Leeuw, Edith, Natacha Borgers, and Astrid Smits. 2004. "Pretesting Questionnaires for Children and Adolescents." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

DeMaio, Theresa, and Ashley Landreth. 2004. "Do Different Cognitive Interview Techniques Produce Different Results?" In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

DeMaio, Theresa, Nancy Mathiowetz, Jennifer Rothgeb, Mary Ellen Beach, and Sharon Durant. 1993. "Protocol for Pretesting Demographic Surveys at the Census Bureau." Washington, DC: U.S. Bureau of the Census.

Dillman, Don, and Cleo Redline. 2004. "Concepts and Procedures for Testing Paper Self-Administered Questionnaires: Cognitive Interview and Field Test Questions." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Draisma, Stasja, and Wil Dijkstra. 2004. "Response Latency and (Para)Linguistic Expressions as Indicators of Response Error." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Ericsson, K. Anders, and Herbert Simon. 1980. "Verbal Reports as Data." *Psychological Review* 87:215–51.

Forsyth, Barbara, Jennifer Rothgeb, and Gordon Willis. 2004. "Does Pretesting Make a Difference?" In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Fowler, Floyd. 2004. "The Case for More Split-Sample Experiments in Developing Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Gerber, Eleanor. 1999. "The View from Anthropology: Ethnography and the Cognitive Interview." In *Cognition and Survey Research*, ed. Monroe Sirken, Douglas Hermann, Ssuan Schechter, Norbert Schwarz, Judith Tanur, and Roger Tourangeau, pp. 217–34. New York: Wiley.

Graesser, Art, Katja Wiemer-Hastings, Peter Wiemer-Hastings, and Roger Kreuz. 2000. "The Gold Standard of Question Quality on Surveys: Experts, Computer Tools, versus Statistical Indices." *Proceedings of the Section on Survey Research Methods of the American Statistical Association* 459–64.

Hansen, Sue Ellen, and Mick P. Couper. 2004. "Usability Testing to Evaluate Computer-Assisted Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Harkness, Janet, Beth-Ellen Pennell, and Alisú Schoua-Glusberg. 2004. "Survey Questionnaire Translation and Assessment." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Hunt, Shelby, Richard Sparkman, Jr., and James Wilcox. 1982. "The Pretest in Survey Research: Issues and Preliminary Findings." *Journal of Marketing Research* 19:269–73.

Jabine, Thomas, Miron Straf, Judith Tanur, and Roger Tourangeau. 1984. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. Washington, DC: National Academy Press.

Kaplowitz, Michael, Frank Lupi and John P. Hoehn. 2004. "Multiple Methods for Developing and Evaluating a Stated-Choice Questionnaire to Value Wetlands." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Katz, Daniel. 1940. "Three Criteria: Knowledge, Conviction, and Significance." *Public Opinion Quarterly* 4:277–84.

Kornhauser, Arthur. 1951. "Constructing Questionnaires and Interview Schedules." In *Research Methods in Social Relations: Part Two*, ed. Marie Jahoda, Morton Deutsch, and Stuart Cook, pp. 423–62. New York: Dryden.

Krosnick, Jon, and Leandre Fabrigar. Forthcoming. *Designing Questionnaires to Measure Attitudes*. New York: Oxford University Press.

Lessler, Judith, and Barbara Forsyth. 1996. "A Coding System for Appraising Questionnaires." In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, ed. Norbert Schwarz and Seymour Sudman, pp. 259–92. San Francisco: Jossey-Bass.

Lessler, Judith, Roger Tourangeau, and William Salter. 1989. "Questionnaire Design Research in the Cognitive Research Laboratory." *Vital and Health Statistics* (Series 6, No. 1; DHHS Publication No. PHS-89-1076). Washington, DC: Government Printing Office.

Loftus, Elizabeth. 1984. "Protocol Analysis of Responses to Survey Recall Questions." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, ed. Thomas Jabine, Miron Straf, Judith Tanur, and Roger Tourangeau, pp. 61–64. Washington, DC: National Academy Press.

Martin, Elizabeth. 2004. "Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Martin, Elizabeth, Susan Schechter, and Clyde Tucker. 1999. "Interagency Collaboration among the Cognitive Laboratories: Past Efforts and Future Opportunities." In *Statistical Policy Working Paper 28: 1998 Seminar on Interagency Coordination and Cooperation*, pp. 359–87. Washington, DC: Federal Committee on Statistical Methodology.

Moser, Claus, and Graham Kalton. 1971. *Survey Methods in Social Investigation*. London: Heinemann.

Moore, Jeffrey, Joanne Pascale, Pat Doyle, Anna Chan, and Julia Klein Griffiths. 2004. "Using Field Experiments to Improve Instrument Design." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Oksenberg, Lois, Charles Cannell, and Graham Kalton. 1991. "New Strategies for Pretesting Survey Questions." *Journal of Official Statistics* 7:349–56.

Presser, Stanley, and Johny Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" *Sociological Methodology* 24:73–104.

Presser, Stanley, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer, eds. 2004. *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley.

Reeve, Bryce, and Louise Mâsse. 2004. "Item Response Theory (IRT) Modeling for Questionnaire Evaluation." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Rothwell, Naomi. 1983. "New Ways of Learning How to Improve Self-Enumerative Questionnaires: A Demonstration Project." Unpublished manuscript, U.S. Bureau of the Census.

———. 1985. "Laboratory and Field Response Research Studies for the 1980 Census of Population in the United States" *Journal of Official Statistics* 1:137–57.

Royston, Patricia, and Deborah Bercini. 1987. "Questionnaire Design Research in a Laboratory Setting: Results of Testing Cancer Risk Factor Questions." *Proceedings of the Survey Research Methods Section of the American Statistical Association* 829–33.

Saris, Willem, William van der Veld, and Irmtraud Gallhofer. 2004. "Development and Improvement of Questionnaires Using Predictions of Reliability and Validity." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Schaeffer, Nora Cate, and Jennifer Dykema. 2004. "A Multiple-Method Approach to Improving the Clarity of Closely Related Concepts." In *Methods for Testing and Evaluating Survey*

Questionnaires, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Schaeffer, Nora Cate, and Stanley Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29:65–88.

Sheatsley, Paul. 1983. "Questionnaire Construction and Item Writing." In *Handbook of Survey Research*, ed. Peter Rossi, James Wright, and Andy Anderson, pp. 195–230. New York: Academic Press.

Sirken, Monroe, Thomas Jabine, Gordon Willis, Elizabeth Martin, and Clyde Tucker, eds. 1999. *A New Agenda for Interdisciplinary Research: Proceedings of the CASM II Seminar*. Hyattsville, MD: National Center for Health Statistics.

Sletto, Raymond. 1950. "Pretesting of Questionnaires." *American Sociological Review* 5:193–200.

Smith, Tom. 2004. "Developing and Evaluating Cross-National Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Stouffer, Samuel, Louis Guttman, Edward Suchman, Paul Lazarsfeld, Shirley Star, and John Clausen. 1950. *Measurement and Prediction*. Princeton, NJ: Princeton University Press.

Sudman, Seymour. 1983. "Applied Sampling." In *Handbook of Survey Research*, ed. Peter Rossi, James Wright, and Andy Anderson, pp. 145–194. New York: Academic Press.

Sudman, Seymour, and Norman Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass.

Sudman, Seymour, Norman Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Tarnai, John, and Danna Moore. 2004. "Methods for Testing and Evaluating Computer-Assisted Questionnaires." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Tourangeau, Roger. 2004. "Experimental Design Considerations for Testing and Evaluating Questionnaires." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Tourangeau, Roger, Lance Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

U. S. Census Bureau. 2003. Census 2000, Summary File 3, Tables P19, PCT13, and PCT14. Summary Tables on Language Use and English Ability: 2000 (PHC-T-20).

Van der Zouwen, Johannes, and Johannes Smit. 2004. "Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question-Answer Sequences." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Willimack, Diane, Lars Lyberg, Jean Martin, Lilli Japec, and Patricia Whitridge. 2004. "Evolution and Adaptation of Questionnaire Development, Evaluation and Testing for Establishment Surveys." In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.

Willis, Gordon. 1994. *Cognitive Interviewing and Questionnaire Design: A Training Manual*. Hyattsville, MD: National Center for Health Statistics

———. 2004. "Cognitive Interviewing Revisited: A Useful Technique, in Theory?" In *Methods for Testing and Evaluating Survey Questionnaires*, ed. Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: Wiley.