2

SURVEY PRETESTING: DO DIFFERENT METHODS PRODUCE DIFFERENT RESULTS?

*Stanley Presser**
*Johnny Blair**

Pretesting is generally agreed to be an indispensable stage in survey questionnaire development, yet we know little about how well different pretesting methods identify particular types of problems. This study compared four pretesting methods using a single questionnaire in repeated trials of each. The four methods were conventional pretests, behavior coding, cognitive interviews, and expert panels. We developed a model-based coding scheme that classified problems as respondent-semantic, respondent-task, interviewer-task, or analysis. On average, expert panels were most productive in the number of problems identified. Conventional pretesting and behavior coding were the only methods to identify significant numbers of interviewer problems. By contrast, expert panels and cognitive interviews were the only methods to diagnose a nontrivial number of analysis problems. Expert panels and behavior coding were more consistent than the other methods in the numbers of problems identified across trials, as well as in their distribution of problem types. From the vantage point of the particular problems

An earlier version of this paper was presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, Florida, May 17, 1992. The paper is based on research funded by the National Center for Health Statistics. We thank Christine Anthony and Beth Webb, who coded the data, Dana Wagner, who supervised the coding, Timothy Triplett, who designed and constructed the data files, and Robert Groves, Howard Schuman, the editor, and the anonymous referees, who provided helpful comments.

*University of Maryland

identified, behavior coding was the most reliable method. Costs of the conventional pretests and behavior coding were about the same, cognitive interviews were somewhat less expensive, and expert panels were considerably cheaper.

Many survey researchers would agree with Warwick and Lininger (1975, p. 161) that "the absolute necessity of pretesting the questionnaire . . . cannot be overemphasized." Nonetheless, Backstrom and Hursh-Cesar (1981, p. 285) hardly exaggerated when they wrote "The pretest is the most misunderstood and abused element of the survey process." Indeed, "to say a question has been pretested, may be saying very little" (Presser 1989, p. 35). This is because we know only a modest amount about how pretest data are collected, how they are evaluated, and how well they identify questionnaire problems.

In the most common form of pretesting a few interviewers each conduct a small number of interviews and then attend a debriefing in which they discuss their experiences. Response distributions from pretest interviews can be analyzed, but our sense is that this is not typically done. Instead, the data from most pretests consist of interviewer reports about their interviews. As a result, conventional pretests reflect respondent reactions mainly via the perceptions of the interviewers.

Various factors may affect the likelihood that interviewers will be sensitive to and convey respondent reactions. These include interviewer experience [Nelson (1985) discusses the advantages and disadvantages of different mixes of experience]; the methods interviewers use to report their observations [DeMaio (1983) describes written questionnaires that may be used to supplement oral debriefings]; and whether or not respondents are informed they are taking part in a pretest [Converse and Presser (1986) distinguish between "participating" and "undeclared" pretests]. Yet there is no escaping the fact that in the conventional pretest, the task of judging respondent reactions is basically a subjective one.

About 25 years ago, Charles Cannell and his associates at the University of Michigan devised a more objective yardstick to assess problems with questionnaires. They constructed a formal code for the interaction that occurs between interviewer and respondent during questionnaire administration (Cannell and Robison 1971; Marquis 1971). Underlying this approach is the assumption that the interaction

created by "problem-free" questions will conform to the norm of the interviewer's reading the question verbatim followed by the respondent's providing an acceptable answer. By listening to a tape-recording of an interview, or monitoring an interview in progress, a coder follows the interviewer-respondent exchange question-by-question. Each time a behavior deviating from the norm occurs—such as the interviewer departing from the exact questionnaire text, or the respondent requesting clarification—a particular "behavior code" is assigned to the question. Questions scoring high on behaviors departing from the norm (e.g., respondent requests to have an item repeated) are assumed to have problems.

Partly because Cannell and his associates became more interested in using behavior codes to study the performance of interviewers than of questionnaires (Cannell et al. 1975 and 1981), and partly because questionnaire testing was not generally considered an area in need of work, this approach to pretesting gained few adherents at the time.[1] Recently, however, research in federal statistical agencies on ways to improve questionnaires has led to renewed interest in using behavior codes as a testing strategy (Cannell et al. 1989; Fowler 1989; Willis 1991).

Also in recent years, the federal statistical establishment has contributed to the development of another method of questionnaire testing. Responding to interest in the application of cognitive psychology to survey research (Jabine et al. 1984), cognitive labs were created first at the National Center for Health Statistics (Royston et al. 1986) and then at the Bureau of Labor Statistics and the Census Bureau. Pretesting in these labs involves the use of methods designed to explore respondent cognitive processes. The procedures vary widely, but most include administering survey questions to a subject, who is asked to "think aloud," reporting everything that comes to mind while arriving at an answer, or to report immediately after answering how the answer was reached. The resulting "concurrent" or "retrospective" think-alouds are analyzed to determine comprehension and the strategies respondents use to come up with answers. Interviewer probes (e.g., "What did you think I meant by . . . ?") may also be used to focus on aspects of questions the researcher suspects may be

[1]A notable exception was work by Jean Morton-Williams in England (Morton-Williams 1979; Morton-Williams and Sykes 1983).

problematic for respondents. (For a taxonomy of the various cognitive methods, see Forsyth and Lessler 1991.)

In some respects the cognitive approach to pretesting implements the much-ignored advice to conventional pretesters to ask respondents questions about the questions. As Selltiz et al. (1959, p. 551) counseled: "A valuable part of the pretest interview is discussion of the questions with respondents after they have answered them. The respondent may be asked what the question meant to him, what difficulties he experienced in replying, what further ideas he had that were not brought out by the question . . . etc."[2]

The three pretesting approaches (conventional, behavior coding, and cognitive) seem very different. Indeed, it is easy to imagine them yielding different conclusions about the same questions. Consider, for example, the item, "Do you own a car?" Cognitive interviews seem better designed than conventional pretests to uncover the diverse interpretations it occasions: "you" can be interpreted as singular, referring only to the respondent, or as plural, encompassing all household members; "own" can convey the legal sense of "title" or the practical sense of "have," including cars bought on credit or leased long term; and "car" can be viewed as including vans and trucks, or seen more narrowly as automobile. By contrast, consider the question "Do you or anyone else in your household own or lease a motor vehicle such as a car, truck, or van?" Behavior coding seems better designed than cognitive interviews to reveal that its length and complexity cause interviewers to misread it.

Conventional pretests might appear to be particularly strong in identifying problems that cause difficulty for the interviewer, whereas the greatest strength of cognitive interviews would seem to be in identifying problems that cause difficulty for the respondent. The focus of behavior coding on both sides of the interviewer-respondent interaction suggests that it might be equally strong at identifying the two kinds of problems. Behavior coding would also appear to be better suited to diagnose respondent uncertainty about a question's meaning than to reveal the meaning respondents attri-

[2]The unscripted character of these follow-up probes places a considerable burden on the interviewer and is probably responsible for the method only occasionally being carried out by conventional pretesters. For an early example that systematically used this approach, see Belson (1981); for a recent innovative variation, see Campanelli et al. (1991a).

bute to questions they do understand—a probable strength of cognitive interviews.

Despite these grounds for believing that some pretest methods are better than others at identifying particular classes of problems, there is not much evidence demonstrating this. Similarly, little is known about other fundamental ways in which pretest methods may vary. Each of the methods can be characterized in terms of reliability (the degree to which the method yields similar conclusions across independent trials), validity (the extent to which it identifies problems that would affect a survey's results), and cost (the amount and kind of resources it requires).

Do different pretest methods vary on these dimensions? To address this question, we tested a single questionnaire in repeated trials of multiple methods. The results provide estimates of the reliability and cost of each method as well as of the kinds of problems each method identifies. As will become apparent, however, our design is less well suited to estimating the validity of the methods.

A few prior studies have compared the findings of different approaches to testing the same questionnaire. Hunt et al. (1982) found that participating pretests conducted by telephone were more successful than those conducted as mall intercepts in identifying errors that had purposely been built into a test questionnaire. Lessler et al. (1989) reported that the problems identified by a conventional undeclared pretest of a Health Interview Survey supplement were fairly similar, though not identical, to those identified in a cognitive laboratory, but they suggested that the latter was the more efficient method as it involved many fewer interviews. Bischoping (1989) compared two conventional pretests with each other and also compared one of them with the results of behavior coding. She found agreement ranged from quite low to high depending on the type of problem considered. Most recently, Campanelli et al. (1991*b*) and Esposito et al. (1991) contrasted three features of prestests used in the redesign of the Current Population Survey: behavior coding, interviewer debriefings, and respondent debriefings that included specially designed probes of earlier answers. They judged the respondent debriefings to be the most helpful method, but concluded that each of the methods had strengths and weaknesses.

Conventional pretests, cognitive interviews, and behavior coding are alike in that they all involve collecting data from members of

the population of interest. The rationale for pretesting is that trying out a questionnaire on respondents will reveal problems that even the most experienced survey researcher cannot diagnose. Although this is an article of faith in the survey research world, it seems worth testing. Thus we added to the original three methods a fourth method involving a review by survey research experts.

1. DATA COLLECTION

1.1. *Test Questionnaire*

Our test questionnaire was composed of five National Health Interview Survey supplements in early stages of development. Supplements were selected to provide a wide range of question types. The topics were Food Knowledge, Dietary Behavior, Medical Care, General Health Knowledge, and Knowledge about AIDS. In order to remain within the budgeted 30-minute interview and to include only questions that would apply to all members of the general population, parts of some sections had to be cut. Care was taken to maintain the internal consistency of each section, but no changes were made to question wordings. The final instrument consisted of 140 questions (some of which were brief items asked as part of a series).

To prevent contamination across trials, the replications of each method were carried out by separate staffs who were instructed not to discuss the project with others. Staff assigned to all treatments were given the identical questionnaire accompanied by a set of questionnaire objectives compiled from National Center for Health Statistics documents about the supplements. All data collection took place at the University of Maryland Survey Research Center during 1991.

1.2. *Conventional Pretests*

Two undeclared pretests, each staffed by four interviewers, were carried out with a random digit dial sample of adults. All eight interviewers had prior pretesting experience. Thirty-five telephone interviews were completed in the first pretest, and 43 in the second. At the conclusion of interviewing, separate debriefings were conducted by two different senior staff members. Interviewers reported on their overall experience and went through the questionnaire item by item.

The debriefings were tape-recorded and the senior staff members prepared summary reports describing the problems identified.

1.3. *Behavior Coding*

We had planned to code interviewer-respondent interaction in all the conventional pretest interviews, but at times there were more interviewers working than trained monitors. As a result 60 percent of the interviews from the first pretest were coded and 70 percent from the second. The interviewers were unaware special monitoring was occurring (though they expected that some of their work would be monitored as part of regular quality control). Because coding was done in real time (not from tapes), the coding scheme had to be kept relatively simple. The codes representing problems were: Interviewer makes a major change in reading in the question; Interviewer has to probe; Respondent asks for clarification or has some other difficulty with the question; and Respondent gives an uncodable answer.

1.4. *Cognitive Interviews*

Face-to-face cognitive interviews were conducted with three sets of 10 to 12 individuals recruited from the local telephone book using gender, age, and education quotas. Respondents were paid $25 to come to the Center. Each set of interviews was conducted by a different interviewer who had previous experience doing cognitive interviewing for questionnaire development. One of the interviewers had a doctorate in cognitive psychology; the other two had no academic training in cognitive psychology. The interviewers used a combination of follow-up probes and concurrent and retrospective think-alouds. All the interviews were tape-recorded and each of the interviewers prepared a summary report describing the problems encountered.

1.5. *Expert Panels*

Two panels of survey research experts were asked to review the questionnaire for problems. One member of each panel was a psychologist by training, one was a specialist in questionnaire design, and one was a general survey methodologist. Panel discussions lasted from two to three hours and were tape-recorded. In addition, the

member of the group who led the discussion prepared a summary report of the panel's points.

2. SUMMARY REPORTS

The instructions about how to prepare the summary reports were the same for all methods: Provide a listing of the problems identified for each question. Yet the flavor of the reports varied across methods. This is because the methods differed in the kinds of data they produced as well as in the perspectives and backgrounds of those who carried them out. The expert panels were distinctive in that they involved no formal data collection,[3] and their staff had the greatest familiarity with research on questionnaires. This combination was reflected in diagnoses such as "The recall period for these items is far too long; either give respondents a diary or ask them only about a few days" and "Covering eating and drinking in a single item [about diet changes] and limiting changes to those made 'for health reasons' make it cognitively more difficult."

By contrast, the conventional pretest reports typically drew on the interviews themselves. For instance, "For question 32 [about the two leading causes of death from disease], respondents always choose only one option instead of the requested two" and "Interviewers had difficulty reading question 36a."

The cognitive reports also contained frequent examples of respondent behaviors, some of which were similar to those from the conventional pretests—for example, "Nine respondents failed to note that the question [number 32] refers to 'two of the following diseases . . .' and reported only one disease." But many of the problems identified in the cognitive reports were presented in an explicit psychological framework—for example, "Subjects often requested that the three statements be reread. This seemed to be a case of it being hard to maintain semantically similar items in working memory, especially [items] as wordy and as similar as these."

The reports varied in the extent to which they gave the basis for a problem's diagnosis, or described the problem in enough detail to allow an informed inference about its cause. In line with the

[3]In fact, a member of one of the panels conducted a single interview with the questionnaire prior to the discussion. Nothing in the summary report, however, makes direct mention of it.

instructions, however, each report provided an indication of what the problem was in every case. We therefore decided to focus on problem types in making comparisons both within and across methods (with the problem being the unit of analysis).

3. PROBLEM CODING

Only the results from the behavior coding lend themselves straightforwardly to quantification, as that method records the frequency of kinds of problematic behavior for each question. We focused on two types of behaviors, those indicating respondent difficulty (including Interviewer has to probe, Respondent asks for clarification or has some other difficulty with the question, and respondent gives an uncodable answer) and those indicating interviewer difficulty (Interviewer makes a major change in reading the question). Of course, deciding on the frequency of each of these behaviors that constitutes a problem requires selection of an arbitrary cutoff. In accordance with Fowler's (1989) recommendation, we judged that a problem was present if it was coded at least 15 percent of the time a question was asked, but we examined the sensitivity of results to using lower (10 percent) and higher (20 percent) cutoffs.

We determined which problems each of the other methods diagnosed on the basis of the summary reports. The reports were organized question by question. Everything considered during the trial was not necessarily included. The chair of an expert panel, for example, may not have included mention of a problem that was raised by one member on which consensus was not reached. Likewise the senior staff member who led a conventional debriefing may not have included a problem mentioned by a single interviewer, if the ensuing discussion (or the senior staff member's judgment) suggested it was idiosyncratic.

Relying on the summary reports (which incorporated authors' judgments), as opposed to coding the transcripts of the debriefings, panel discussions, and cognitive interviews, accords with typical survey practice, in which someone judges what problems have been identified on the basis of conducting or observing the debriefing, discussion, or cognitive interviews. (Tape recordings are usually not made, and, in cases where they are available, frequently not consulted because of time constraints.) In making these kinds of judg-

ments, survey researchers are sometimes confronted with conflicting opinions or meager evidence. Although in many instances all the pretest participants will agree that a question has a problem, in others there will be disagreement. Sometimes the occurrence of a problem on just a single occasion will be convincing; at other times even a few occurrences may be judged an anomaly. The researcher must exercise his or her judgment in such cases.

All the summary reports were read independently by two trained coders who carried out the interrelated tasks of identifying the total number of different problems mentioned in a report and assigning each one to a problem type. This work was guided by a coding scheme that we devised. The scheme was developed partly inductively (by grouping problems identified in pretests of three other surveys), but mainly by modeling the stages in the survey process where problems caused by questions can occur. As may be seen in Figure 1, the model draws heavily on earlier models of the response process (e.g., Cannell et al. 1981).

The survey process begins with the specification by the investigator of a subject and task for each question. In the question "How many times have you visited a physician in the last six months?" the subject is physician visits during a specified time and the task is to recall and report how many of them occurred.

In the second stage, the interviewer asks the question of the respondent. This is followed by the respondent engaging in the semantic work necessary to interpret the subject and task.

In the next stage, respondents carry out the task called for by their understanding of the question. This involves retrieving relevant data from memory, formulating a response, and providing an answer. Then the interviewer records a response. Finally, the analyst uses the data.

Problems with the question may cause difficulty at any of these stages. For example, question problems may affect whether the interviewer reads the item exactly as specified; whether an item is understood and how; whether the respondent can retrieve or report an answer; whether there are discrepancies between the respondent's answer and what the interviewer records; and whether the data analysis is beset by complications such as response sets.

We arrived at four major categories of questionnaire problems based on this model: one affecting the interviewer, two affecting the

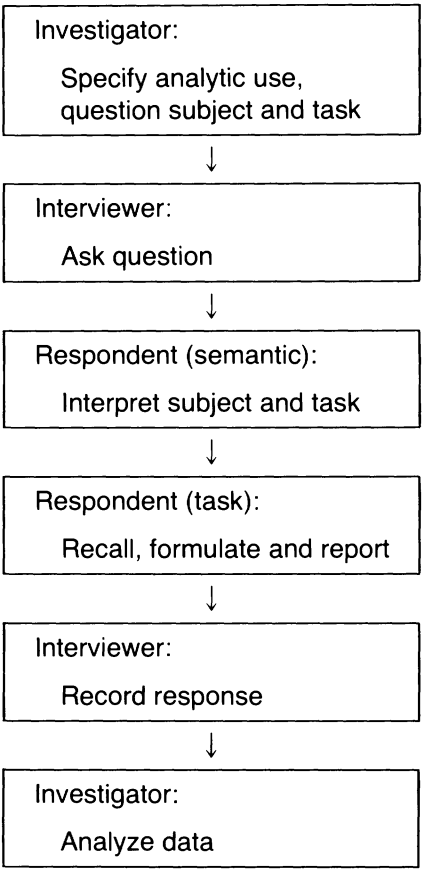

FIGURE 1. Overview of model of the survey process.

respondent (semantic and task), and one affecting the data analyst. [We initially used two different semantic categories—labeled I and II below—in an attempt to distinguish difficulties that affected the respondent's arriving at enough of an understanding of the question to attempt to answer it (things that interfered with discerning meaning) from those that affected the particular meaning the respondent arrived at. However, the coders were unable to make this distinction reliably, so we abandoned it and combined the two categories in our analyses.]

The coder had to decide whether each statement in the summary reports indicated that an item would pose a problem for some

respondents, interviewers, or analysis needs. The statement might reflect an actual problem that was observed, or (particularly for the expert panels) the report preparer's judgment about the likelihood of a problem. To classify problems, the coder answered the following questions in order, stopping if a "yes" answer was obtained:

1 Does the statement mean the respondent has any difficulty coming to an understanding of what the question means? YES = Semantic I;
2 Does the statement mean the respondent has any difficulty remembering the question? YES = Semantic I;
3 Does the statement mean the respondent has any difficulty understanding the meaning of particular words or concepts in the question? YES = Semantic II;
4 Does the statement mean different respondents have different understandings of what the question refers to? YES = Semantic II;
5 Does the statement mean the respondent has any difficulty recalling, formulating, or reporting an answer? YES = Respondent Task;
6 Does the statement mean the interviewer has any problem reading the question or recording the answer? YES = Interviewer;
7 Does the statement mean the analyst has any problem using the data? YES = Analysis.

The phrase "has a problem (or difficulty)" was to be interpreted as "has, or would have, a problem (or difficulty)." If all seven questions were answered "no," the statement was not coded as identifying a problem.

Thus each problem was coded in only a single category even if it could cause several kinds of difficulties, giving priority first to respondent difficulties, then to interviewer difficulties, and lastly to potential analysis difficulties. In addition, we assumed semantic difficulties were more basic than task difficulties.[4] For example, the state-

[4]This set of priorities, which affects only the coding of those problems described as causing difficulty at more than one stage in the survey process (see Figure 1), was an attempt to reflect the order in which problems actually occur. Thus a problem that affects both the semantic and task stages causes the semantic difficulties first. In retrospect, we realized that interviewer problems affecting

ment "This item may be hard for some respondents, since the [answer] may vary depending on whether the trip was from home or the office" indicates both a semantic and a task problem, but would be coded only in the semantic category as the problem respondents might have doing the task (determining how long the trip takes) would be a consequence of their problem understanding what the question means (whether it refers to trips from home or office).

In the "trip" example, since the same aspect of the question could cause two kinds of problems only the more basic problem would be coded. However, if a statement identified two problem types stemming from different aspects of a question, both would be coded. Thus the statement "The word 'inherent' is too difficult for some respondents and the agree-disagree format may produce acquiescence response set" would be coded in both the semantic and analysis categories. The detailed coding frame (which made finer distinctions within each of the four major categories) is included in the appendix to this chapter.[5]

Our analysis assumes that all the things identified as problems by the different methods really do constitute flaws in the questions. In terms of the behavior codes, it seems reasonable to believe that questions associated more than 15 percent of the time with actions like requests for clarification are in fact problematic. In terms of the other methods, our belief that the problems identifed are actually question defects rests on three grounds: (1) the report writers, who are experienced survey professionals, judged them to be problems; (2) the coders, in applying the coding scheme (based on a model of the stages at which questions can lead to problems), identified them

how the question is read (as opposed to how the answer is recorded) should have been given priority over the two respondent problem types, as such interviewer problems occur first. Our review of problems coded in the semantic and task categories, however, did not reveal any that were described as also causing an interviewer difficulty. Thus our results are unaffected by this decision.

[5]Of all the problems identified by the coders, 72 percent were identified by both. In the remaining cases, the coders either disagreed that a particular statement identified a problem with a question or disagreed about the number of different problems the statement identified. When the coders agreed there was a problem, they made the same assignment from among the four major problem types 89 percent of the time. Cohen's kappa for the semantic code was 0.82; for the task code, 0.80; for the interviewer code, 0.89; and for the analysis code, 0.76. All between-coder differences were reconciled by the supervisor who had trained the coders.

as problems; and (3) our own review made us feel they were problems (expressed differently, if we were undertaking a revision of the questionnaire, we would address them).

Of course, problems vary in their severity. Some will have a major effect on responses, others a minor effect, and still others no effect at all. Thus, ideally, we would like to know whether the problems identified by the different methods vary in their seriousness. Unfortunately, only an experimental comparison can establish that a particular aspect of a question will affect responses and by how much. Thus our present design does not allow us to make judgments about problem seriousness.

4. RESULTS

4.1. *Overall Productivity*

Table 1 displays the number of problems identified in each of the nine trials as well as the distribution of the types of these problems. Averaging across trials, expert panels were the most productive method, identifying roughly 160 problems, compared to about 90 for conventional pretests, cognitive interviews, and behavior coding.[6] Averaging across trials, however, conceals considerable variation for two of the methods. One of the conventional pretests reported the smallest number of problems of the nine trials (27), the other, the next to largest number (154). Cognitive interviews also showed substantial, though less extreme, variability in productivity, trial 3 reporting over twice as many problems as trial 1. By contrast, the expert panels were fairly similar in this regard, and the behavior coding trials were the most alike.

4.2. *Problem Types*

Behavior coding and expert panels were also the most reliable methods in the distribution of the kinds of problems identified. In both behavior coding trials, about 85 percent of the problems reflected

[6]In our implementation, behavior coding could be credited with only a single problem per question for each of its types (respondent and interviewer), whereas the other methods could identify two or more different problems of the same type per item.

TABLE 1
Number and Types of Problems Identified by Trial

	Total Number of Problems	Respondent Semantic	Respondent Task	Respondent Behavior	Interviewer	Analysis	Total Percentage
Conventional 1	154	41.6%	44.2	—	12.3	1.9	100%
Conventional 2	27	63.0%	14.8	—	22.2	0	100%
Cognitive 1	54	70.4%	13.0	—	0	16.7	100%
Cognitive 2	92	62.0%	20.7	—	0	17.4	100%
Cognitive 3	138	76.1%	21.7	—	1.4	0.7	100%
Expert 1	182	45.1%	34.6	—	1.1	19.2	100%
Expert 2	140	55.7%	27.9	—	1.4	15.0	100%
Behavior coding 1	89	—	—	88.8%	11.2	—	100%
Behavior coding 2	102	—	—	82.4%	17.6	—	100%

	x^2_{LR}	d.f.	p
Difference between 9 trials	162.11	19	0.001
Difference between methods (collapsing across trials of a method)	113.41	6	0.001
Difference between conventional trials	10.81	3	0.01
Difference between cognitive trials	32.46	6	0.001
Difference between expert trials	3.84	3	n.s.
Difference between behavior trials	1.59	1	n.s.

respondent difficulties and 15 percent, interviewer difficulties.[7] For each of the expert panels, roughly half the reported problems were respondent-semantic ones, a third respondent-task, and a sixth analysis.

Although cognitive interviews and conventional pretests were more variable across trials in the types of problems identified, each was nonetheless distinctive. The lion's share of problems from all three sets of cognitive interviews were semantic ones, with virtually no interviewer problems reported. By contrast, conventional pretesting (in both trials) was the only method other than behavior coding that reported a significant proportion of interviewer problems, but it reported virtually no analysis problems.

If one treats the results as a 9 by 5 contingency table with structural zeros in cells that can have no entries (because the method could not assign a particular problem type), the difference among the nine trials in the types of problems identified was highly significant ($X^2 = 162.11$). The overall chi-square can be partitioned into one component due to differences between the methods (collapsing across trials of each method) plus four components representing the differences between trials of the same method. Partitioning in this manner, a little more than two-thirds of the original chi-square stems from differences between methods, with the remainder due to the difference between the conventional trials and the variability among the cognitive trials. Neither of the other components, one representing the difference between the expert panels, the other between the behavior codings, was statistically significant.[8]

Of course, similar distributions of problem types across trials of the same method (as for the expert panels and behavior codings) could have occurred without the trials agreeing on which questions had problems. Trials could have identified the same kinds of problems but diagnosed different questions as having them. Conversely, different distributions of problem types could have occurred even

[7]The results are fairly similar if the behavior level that defines a problem is set at either 0.10 or 0.20 (instead of 0.15). At the lower cutoff, 83 percent and 74 percent of the problems in trials 1 and 2, respectively, were in the respondent category; the corresponding percentages at the higher cutoff were 90 and 86. As would be expected, using 0.10 increases the total number of problems identified, to 109 and 142, and using 0.20 leads to a decrease, to 74 and 85.

[8]Strictly speaking, significance tests are not justified, as the data do not come from a probability sample. We use them for heuristic purposes.

though the trials identified many of the exact same problems. One trial could have identified all the problems found by another and also identified an additional set that changed the distribution.

4.3. *Overlap Between Trials*

To assess the degree to which different trials in fact diagnosed the same problems, we listed all the problems identifed in the nine trials, deleting duplicates.[9] This yielded a total of 408 different problems.[10] Each trial was then coded 0 or 1 for each of the 408 problems, depending on whether it identified the problem. We thus have a data set with cases (representing problems) and variables (coded found/not found) representing trials.

Table 2 shows the relationships between each pair of these trial variables for the three methods for which the full set of respondent, interviewer, and analysis problem types could be coded. We expected that the association between trials of the same method would be higher than that between trials of different methods. The results provide, at best, weak support for this idea.

Overall, the cognitive trials were somewhat more strongly related to each other than to trials of other methods, but there were notable exceptions (e.g., the first cognitive trial produced a stronger relation to the second conventional trial than to either of the other cognitive trials). On average, the two expert trials showed only marginally higher within-method than between-method associations, and there were multiple exceptions (e.g., both expert trials were more strongly associated with the second cognitive trial than they were with each other). Moreover, the two conventional trials were unrelated to each other, although clearly related to many of the trials of other methods. (The significant negative associations between the

[9]To create this list, we first compared all the problems reported in the conventional, cognitive, and expert trials and deleted all duplicates. We then added (1) interviewer problems from either behavior coding trial on questions not identified as having an interviewer problem in any of the seven other trials and (2) respondent problems diagnosed by either behavior coding trial for questions not identified as either respondent-semantic or respondent-task in another trial.

[10]All but two of the 140 questions were diagnosed as having at least one problem, with 111 having more than one problem.

TABLE 2
Overlap in Problem Identification (Yule's Q) Among the Trials of the Three Methods for the Complete Set of Problems

	Conventional 2	Cognitive 1	Cognitive 2	Cognitive 3	Expert 1	Expert 2
Conventional 1	**0.07**	−0.50***	−0.33**	0.04	−0.01	0.10
Conventional 2		0.77***	0.20	0.08	0.45**	0.30
Cognitive 1			**0.60*** **	**0.51*** **	0.41**	0.07
Cognitive 2				**0.37*** **	0.45***	0.48***
Cognitive 3					0.21*	0.44***
Expert 1						**0.37*** **

	Median Within Method	Median Between Methods
Conventional	0.07	0.09
Cognitive	0.51	0.20
Expert	0.37	0.36

Note: Coefficients in boldface are within method.

$n = 408$

***$p < 0.001$

**$p < 0.01$

*$p < 0.05$

first conventional trial and two of the cognitive trials are especially puzzling.)

This kind of analysis can be carried out for the behavior trials only by ignoring the distinction between semantic and task problems and deleting analysis problems. Combining the two respondent categories and deleting the analysis category reduces the total number of problems to 285.[11] For this set of 285 problems, the two behavior trials were much more strongly associated (0.89) than either was with trials of other methods (on average, 0.42).

In order to examine whether this much higher reliability for behavior coding was due to its identifying only two (as opposed to four) different kinds of problems, we redid the analysis for the other methods using the reduced set of 285 problems. Table 3 shows that although many of the associations increased substantially after excluding the analysis category and collapsing the semantic and task categories, this increase occurred both within and between methods. As a result, there was again almost as strong a relation between trials of different methods as between trials of the same method. Thus compared to the other methods, behavior coding was indeed distinctive on this measure of reliability.[12]

4.4. *Costs*

Table 4 displays the actual direct costs of our conventional, cognitive, and behavior coding trials. These are probably reasonable estimates of the costs of such tests in many survey organizations (though the cognitive figures assume the availability of staff members already trained in conducting such interviews). This is not true of our costs for the expert panels, which consisted mainly of a token honorarium to the chairs. As the usual survey will not have our good fortune in persuading experienced reseachers to donate their time, we have estimated the cost of this approach assuming the panelists were to be

[11]Deleting the analysis category reduces the number of problems by 75. Ignoring the distinction between the semantic and task categories accounts for the remaining reduction of 48 problems. (There were 48 questions with a single *semantic* problem and a single *task* problem. In the reduced set, these 96 problems were treated as 48 *respondent* problems.)

[12]This is equally true if the behavior level that defines a problem is set at 0.10 or 0.20.

TABLE 3
Overlap in Problem Identification (Yule's Q) Among the Trials of All Four Methods for the Reduced Set of Problems

	Conventional 2	Cognitive 1	Cognitive 2	Cognitive 3	Expert 1	Expert 2	Behavior 1	Behavior 2
Conventional 1	**0.71*** **	0.93***	0.41***	−0.07	0.61***	0.60***	0.55***	0.49***
Conventional 2		0.81***	0.60***	−0.24	0.41*	0.51**	0.23	0.12
Cognitive 1			**0.85*** **	**0.65*** **	0.86***	0.93***	0.42**	0.41**
Cognitive 2				**0.40** **	0.65***	0.91***	0.51***	0.48***
Cognitive 3					0.23*	0.36**	0.32**	0.09
Expert 1						**0.64*** **	0.47***	0.40***
Expert 2							0.74***	0.70***
Behavior 1								**0.89*** **

	Median Within Method	Median Between Methods
Conventional	0.71	0.50
Cognitive	0.65	0.45
Expert	0.64	0.60
Behavior	0.89	0.44

Note: Coefficients in boldface are within method.

$n = 285$

$^{***}p < 0.001$

$^{**}p < 0.01$

$^{*}p < 0.05$

TABLE 4
Costs per Trial

Conventional Pretest	Trial 1	Trial 2	
Training	$234	$233	
Data collection	1,664	1,658	
Debriefing	544	698	
Total	$2,442	$2,589	
Behavior Coding	**Trial 1**	**Trial 2**	
Conventional less debriefing	$1,898	$1,891	
Training	125	125	
Monitoring	190	151	
Tallying	326	454	
Total	$2,539	$2,621	
Cognitive Interviews	**Trial 1**	**Trial 2**	**Trial 3**
Recruiting	$665	$665	$603
Interviewing and report	1,025	1,025	1,025
Subject fees	250	300	250
Total	$1,940	$1,990	$1,878
Expert Panel	**Trial 1**	**Trial 2**	
Discussion leader/report	$440	$440	
Participants ($55/hr × 3 × 5)*	825	825	
Total	$1,265	$1,265	

*Based on a $435 per day government consulting rate.

paid $55 an hour (based on the standard $435 per day consultant's rate used in government grants and contracts in 1991).

Table 4 reveals that the costs of the conventional and behavior coding approaches were essentially the same. The cognitive interviews cost about 20 percent less and the expect panels roughly 50 percent less. It should also be noted that expert panels required somewhat less time than the other methods.

5. DISCUSSION

Of the four methods we tested, behavior coding was the only one with a clearly specified set of objective rules. It is therefore not

surprising that it was the most reliable. It is less clear why conventional pretests were the least reliable. One possibility is that differences in the samples interviewed in the two replications led to very different experiences with the questions. If this had been true, we might expect the results from the behavior trials (based on subsets of these same samples) to differ as well, but this is not the case.

Bischoping (1989) found considerably higher reliability across two conventional debriefings that used about twice as many interviewers (six and nine) and a questionnaire about half as long (60 questions). This suggests that having fewer questions on which to focus and more interviewers to do so may increase reliability.

Despite being somewhat more reliable than conventional pretests, cognitive interviews and expert panels still showed almost as much between-method as within-method consistency. A possible cause of the variation among the cognitive trials is the differences in background of the interviewers, trials 1 and 2 having been conducted by individuals with no academic training in cognitive psychology and trial 3 by someone with a doctorate in the field.[13] If this were a key factor, we would expect trials 1 and 2 to be more strongly associated than either was with trial 3. In fact, the association between trials 1 and 2 (0.60) was considerably stronger than that between trials 2 and 3 (0.37) but not too different from that between 1 and 3 (0.51).

Although the other methods were less reliable than behavior coding, its high reliability comes at a price, as behavior coding (in its usual implementation) provides no information about the cause of a problem. The other methods were more apt to identify the cause of a problem and sometimes to offer a solution as well. In addition, behavior coding could not identify analysis problems or distinguish between respondent semantic as opposed to task difficulties.

In principle, the other methods could have diagnosed all four problem types, but the conventional pretests identified virtually no analysis problems and the cognitive interviews and expert panels yielded almost no interviewer problems. Overall, however, the expert panel was the most cost effective method, being both least ex-

[13]It was the realization, after the first two trials were done, that such formal training might have an impact on the results of the cognitive method that led us to conduct the third trial (in contrast to the two trials of all the other methods).

pensive and most productive. This argues for routinely subjecting questionnaire drafts to a peer review process, in the same way that reports and analyses are peer reviewed.

6. CONCLUSION

Can we generalize from the differences found between pretesting methods in their costs, the numbers and types of problems they identifed, and the reliability (or consistency) with which they performed? Surveys (as well as survey organizations) vary in the way they implement pretesting methods. Conventional pretests, for instance, sometimes involve a handful of interviews, other times hundreds. We might have drawn different inferences about the methods by varying features such as their scale.[14] We believe, however, that our implementation accords with modal survey practice.[15] Nonetheless, we need to be cautious about generalizing from our results, especially as they are based on only two or three trials of each method.[16]

Our analysis has also assumed that the summary reports (upon which the results are based) contained an accurate reflection of the problems diagnosed by the methods. A logical next step is to examine this assumption by coding the problems identified on the tape recordings of the debriefing sessions, cognitive interviews, and expert panel discussions and comparing them to the codes based on the summary reports. (This is, of course, not an issue for behavior coding.)

Future research also needs to assess whether the problems identified by different methods vary in seriousness. Experimental tests com-

[14]We can examine this possibility for the behavior coding trials by randomly dividing each in half. Reducing the number of interviews in this way does not appreciably alter our results. For example, the average number of problems identified by the four smaller trials is 92 (compared to 96 for the two larger trials) and the average reliability across the six pairs of smaller trials is 0.82 (compared to 0.89 for the two larger trials).

[15]Among academic survey research organizations that conduct cognitive interviewing to develop or test questionnaires, the median for the smallest number of such interviews ever conducted on a project is 6 and the median for the largest number ever conducted is 31 (Blair and Presser 1993).

[16]In addition, with one exception, we did not randomly assign staff to the different trials of a method, preferring instead to ensure similarity in the mix of interviewer experience on the conventional pretests and the nature of survey expertise on the expert panels. We did, however, randomly assign respondents to the cognitive trials that were carried out at the same time, as well as to the conventional pretest trials.

paring the original items with revised items that correct the identified problems would reveal this by indicating the extent to which responses are affected by particular problems. In some cases, however, the judgment that a revised item was the better one would depend on knowing the purpose of the item or why it was being asked. For our Health Interview Survey supplements this information was frequently unavailable in the detail needed, and we believe that this is true for many surveys. It is unusual for a survey to have comprehensively specified question-by-question objectives. (This may be partly due to the fact that items are frequently supposed to serve multiple purposes.) Yet without a clear understanding of investigator aims, it can be difficult to assess whether responses to an item reflect those the investigator intends.

Insufficiently specified goals may also be partly responsible for the limitations of conventional pretesting. The objective of most conventional pretests is "to discover problems in the questionnaire," with little further elaboration. The lack of specificity in such a goal may contribute to the method's unreliability.

The recent increase in attention to pretesting issues should contribute to solving both problems. As greater resources are devoted to pretesting, it will become more difficult to avoid specifying item objectives in detail, and it will become easier to specify testing goals in detail. Once testing goals are more clearly thought out, we think it likely that new testing methods, combining features of the different methods we have experimented with, will be devised.

APPENDIX: DETAILED CODING FRAME

NOTE: The coder was routed to a section of this code based on her answers to the questions that appear on p. 84. (The examples in the code do *not* come from the questionnaire used in the pretest trials.)

I. *Semantic—Problems affecting how readily the question is understood or remembered due to:*

A. *Amount of information*

Contains too much text or too many response categories to be retained or understood ("information overload").

Question example: The Census Bureau is required by Congress to count the entire population once every ten years. Census tries to get information about each

person in the country either by sending them a form to complete and mail, or by sending someone to their home to complete the form. Each decade it is becoming more difficult and costly for the Census Bureau to locate and collect information for each person. The Social Security Administration already has in its files information about Census items like date of birth, sex, and place of residence for many people. Some people have suggested that this information be given to the Census Bureau for use in the population census in order to reduce the cost of the Census and the number and length of forms that people have to fill out. How would you feel about the Social Security Administration giving its information about you to the Census Bureau for use in the population census? Would you favor it strongly, favor it somewhat, oppose it somewhat, or oppose it strongly?

Report examples: "This was so long respondents didn't know what I was talking about."
"All the irrelevant information will confuse respondents."
"My respondents had a hard time remembering this."

B. *Structure/organization*

Words or ideas are structured or organized unclearly.

Question Example: Before you got married, how long did you live in Maryland after you graduated from college?

Report Examples: "Respondents asked for this to be repeated, but still weren't sure what they were being asked."
"This may be hard for some respondents to sort out."

C. *Flow or relation between questions*

Question's intelligibility affected by an earlier question or questions (lacks needed transition).

Question Example:

How satisfied are you with police protection in your neighborhood?

How satisfied are you with garbage collection in your neighborhood?

How satisfied are you with the schools in your neighborhood?

How satisfied are you with the grocery stores in the neighborhood where you work?

Report Examples: "You need a transition to alert the respondent that you've switched to their work neighborhood."
"My respondents answered the last question in terms of their residential neighborhood."
"One respondent asked which neighborhood the last question was about."

II. *Semantic—Problems affecting the way the question (or some part of it) is understood due to:*

A. *Boundary lines*

Respondents differ on what the question includes or excludes or are uncertain what the question refers to.

Question Example: How long have you lived in College Park?

Report Examples: "Some respondents who have moved in and out of College Park may count the total time, while others may count only the most recent time."
"It's unclear how students who live here part of the year and with their parents part of the year should answer this."
"Is living in a dorm counted as living here, or only living in a house or apartment? Needs to be clarified."

B. *Insufficient knowledge*

1. Technical term is not understood.

Question example: Do you think that adults should be able to use PCP without any legal penalty?

Report example: "My respondents asked me what PCP was."

2. Common term is not understood (e.g., used in an unusual way).

Question example: For each of the following items, please tell me if your household separates them from your regular garbage so that they can be recycled. First, how about aluminum cans?

Report examples: "Respondents weren't sure what we meant by *regular* garbage."
"Most people don't use the term *regular* to describe some part of their garbage."

C. *Multiple subjects*

A single question asks about more than one subject, each of which could be answered differently ("double barreled").

Question example: Do you think the job the police and courts are doing is excellent, good, fair or poor?

Report examples: "This question asks about two things, police and courts."
"This question is double-barrelled."
"Respondents may have different feelings about the police than the courts."
"Some respondents said the police were doing a good job, but the courts were doing a poor one."

III. *Respondent task—Problems for the respondent retrieving information or formulating or reporting a response*

A. *Recall/response formation is:*

1. Difficult—The level of response detail, demand on memory, or some other feature of the task is too difficult.

Question example: How many times did you go to the movies in the past 12 months?

Report examples: "People who go to a lot of movies will have a hard time with this."
"Respondents will have to estimate; the recall task is too hard."
"People took a lot of time to answer and some said they guessed."

2. Impossible—Information requested is not known.

Question example: "How many miles did you drive in the last year?"

Report examples: "None of my respondents knew this."
"This task cannot be done reliably."

3. Redundant—Answer has (or seems to have) been given to an earlier item.

Question example: "How many times did you start your car's engine yesterday?"
"How many times did you turn off your car's engine yesterday?"

Report examples: "Some of my respondents said I already told you that."
"The second question is redundant."

4. Resisted by respondent—Makes an assumption that is inappropriate or not sensible.

Question example: "How often do you drive over the speed limit on your way to work?"

Report examples: "My respondents who don't drive to work did not know how to answer this."
"This question assumes that people drive to work."

B. *Report affected by:*

1(a) Overlapping response categories

Question example: "What was your total household income from all sources last year? Was it less than $10,000, $10,000 to $25,000, $25,000 to $50,000, or more than $50,000?

Report examples: "My respondent made $25,000 and didn't know what category to use."
"The second and third categories overlap."

1(b) Response categories being insufficient (category is missing)

Question example: "Did you take these courses for professional development or out of personal interest?"

Report examples: "Some of my respondents said both."
"There needs to be a category for people who took courses for both reasons."

1(c). Response categories making too fine a distinction

Question example: I'd like you to rate George Bush using this feeling thermometer. You may use any number from 0 to 100. Ratings between 50 and 100 mean that you feel favorable and warm toward Bush. Ratings between 0 and 50 mean you don't feel favorable. How would you rate your feeling toward Mr. Bush?

Report examples: "My respondents had a lot of difficulty deciding exactly where on the scale they were."
"People can't reliably make use of a 101 point scale."

1(d) Response categories not appropriate to question

Question example: Do you drive to work? NO YES CARPOOL

Report examples: "The third category is an answer to a different question."
"People won't mention carpooling; they'll just say yes or no."

2. *Sensitivity*

Item requires admitting ignorance, undesirable behavior, or something else that leads to discomfort.

Question example: How many sexual partners did you have last year?

Report examples: "Some respondents said they didn't want to answer this."
"Some answers won't be accurate because of the sensitive nature of this."
"Many people will be uncomfortable answering this."

IV. *Interviewer—Problems for the interviewer reading the question or recording the response*

A. *Procedural*

Unclear how the question is supposed to be asked.

Question example: What do you think is the most important cause of crime?
1. Drugs 2. Poverty 3. Guns 4. Criminal justice system

Report example: "I didn't know if I was supposed to read the categories."

B. *Reading problem*

Caused by length, awkward syntax, pronunciation, etc.

Question example: The Social Security Administration has information in its files about Census items like date of birth and sex for nearly everyone. Would you favor or oppose this information being given to the Census Bureau for use in the Census?

Report example: "Interviewers felt that saying 'sex for nearly everyone' sounded silly, and made them stumble at that point.

C. *Coding answers to an open question*

Question example: "Why do you think Jackson should run for President?"
1. Position on the issues 2. Experience 3. Electability 4. Not a Washington Insider

Report examples: "It was hard to fit my respondents' answers into these categories."
"This kind of coding will be difficult for interviewers."

V. *Analysis—Problems for the data analyst because:*

A. *Question is answered the same by all respondents (no variation).*

Question example: "Have you ever been to Tibet?"

Report examples: "All my respondents said no."
"This question is unlikely to produce any variation."

B. *Responses are affected by:*

1. Question suggesting answers or being biased loaded or unbalanced

Question example: How important do you think it is to recycle: extremely important, very important, somewhat important, or not too important?

Report example: "This seems biased towards the importance of recycling."

2. Agree-disagree format ["acquiescence"]

Question example: "President Bush is doing all he can to end the recession." Would you strongly agree, agree, disagree, or strongly disagree?

Report examples: "This question is likely to have an acquiescence bias."
"This may produce bias; a forced-choice item would be better."

3. Order of response categories

Question example: Which of these do you value most in a child: independence, creativity, obedience, cleanliness, self-control, or intelligence?

Report examples: "People may tend to choose the first or last mentioned categories."
"There may be a primacy or recency effect."

REFERENCES

Backstrom, Charles H., and Gerald Hursh-Cesar. 1981. *Survey Research.* 2d ed. New York: Wiley.

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

Bischoping, Katherine. 1989. "An Evaluation of Interviewer Debriefing in Sur-

vey Pretests." In *New Techniques for Pretesting Survey Questions*, by C. Cannell, L. Oksenberg, F. Fowler, G. Kalton, and K. Bischoping, ch. 2. Ann Arbor, Mich.: Survey Research Center.

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

Campanelli, Pamela C., Elizabeth A. Martin, and Jennifer M. Rothgeb. 1991*a*. "The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data." *The Statistician* 40:253–64.

Campanelli, Pamela C., Jennifer M. Rothgeb, James L. Esposito, and Anne E. Polivka. 1991*b*. "Methodologies for Evaluating Survey Questions: An Illustration from a CPS CATI/RDD Test." Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix.

Cannell, Charles F., Sally Lawson, and Doris L. Hausser. 1975. *A Technique for Evaluating Interviewer Performance*. Ann Arbor, Mich.: Institute for Social Research.

Cannell, Charles F., Peter Miller, and Lois Oksenberg. 1981. "Research on Interviewing Techniques." In *Sociological Methodology 1981,* edited by S. Leinhardt, ch. 11. San Francisco: Jossey-Bass.

Cannell, Charles, Lois Oksenberg, Floyd J. Fowler, Graham Kalton, and Katherine Bischoping. 1989. *New Techniques for Pretesting Survey Questions*. Ann Arbor, Mich.: Survey Research Center.

Cannell, Charles, and Sally Robison. 1971. "Analysis of Individual Questions." In *Working Papers on Survey Research in Poverty Areas,* edited by J. Lansing, S. Withey, and A. Wolfe, ch. 11. Ann Arbor, Mich.: Institute for Social Research.

Converse, Jean, and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage.

DeMaio, Theresa, ed. 1983. *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10. Washington: Office of Management and Budget.

Esposito, James L., Pamela C. Campanelli, Jennifer M. Rothgeb, and Anne E. Polivka. 1991. "Determining Which Questions Are Best: Methodologies for Evaluating Survey Questions. In *Proceedings of the Section on Survey Research Methods,* pp. 46–55. Alexandria, Va.: American Statistical Association.

Forsyth, Barbara H., and Judith T. Lessler. 1991. "Cognitive Laboratory Methods: A Taxonomy." In *Measurement Errors in Surveys,* edited by P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman. New York: Wiley.

Fowler, Floyd. J., Jr. 1989. "Coding Behavior in Pretests to Identify Unclear Questions." In *Health Survey Research Methods: Conference Proceedings,* 9–12. Rockville, Md.: National Center for Health Services Research.

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

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

Lessler, Judith. T., Roger Tourangeau, and William Salter, 1989. "Questionnaire Design in the Cognitive Research Laboratory," *Vital and Health Statistics,* series 6, no. 1.

Marquis, Kent. 1971. "Purpose and Procedure of the Tape Recording Analysis." In *Working Papers on Survey Research in Poverty Areas,* edited by J. Lansing, S. Withey, and A. Wolfe, ch. 10. Ann Arbor, Mich.: Institute for Social Research.

Morton-Williams, Jean. 1979. "The Use of Verbal Interaction Coding for Evaluating a Questionnaire," *Quality and Quantity* 13:59–77.

Morton-Williams, Jean, and Wendy Sykes. 1983. "A Study of Question Failure Through the Use of Interaction Coding." *Bulletin of the International Statistical Institute* 1:479–94.

Nelson, Dawn. 1985. "Informal Testing as a Means of Questionnaire Development." *Journal of Official Statistics.* 1:179–88.

Presser, Stanley. 1989. "Pretesting: A Neglected Aspect of Survey Research." In *Health Survey Research Methods: Conference Proceedings,* pp. 35–38. Rockville, Md.: National Center for Health Services Research.

Royston, Patricia, Deborah Bercini, and Monroe Sirken. 1986. "Questionnaire Design Research Laboratory." In *Proceedings of the Section on Survey Research Methods,* pp. 703–707. Washington: American Statistical Association.

Selltiz, Claire, Marie Jahoda, Morton Deutsch, and Stuart W. Cook. 1959. *Research Methods in Social Relations.* New York: Holt.

Warwick, Donald, and Charles Lininger. 1975. *The Sample Survey: Theory and Practice,* New York: McGraw-Hill.

Willis, Gordon B. 1991. "The Use of Behavior Coding to Evaluate a Draft Health-Survey Questionnaire." Paper presented at the annual meeting of the American Association for Public Opinion Research, Phoenix.