Journal of Official Statistics
Vol. 7. No. 3. 1991. pp. 349-365
© Statistics Sweden

New Strategies for Pretesting Survey Questions

Lois Oksenberg, Charles Cannell, and Graham Kalton[1]

Abstract: An experimental study has been conducted to evaluate some new techniques for improving the effectiveness of standard pretesting in identifying problems with questions in survey questionnaires. This paper reports the results obtained from the use of behavior coding and of special probe questions. Coding interview behavior was found to uncover a sizeable number of question problems that would go undetected with the usual pretesting procedures. Special probes were found to be valuable for identifying comprehension problems, but the probes employed for other types of problems were found to be less useful. The addition of these new techniques to standard pretesting leads to a more effective and systematic evaluation of a survey questionnaire without an appreciable effect on the cost or the complexity of the pretest.

Key words: Behavior coding; probe questions; pilot surveys; pretests; questionnaire design.

1. Introduction

The essence of survey research is the collection of information using a standardized questionnaire. The construction of a survey questionnaire frequently involves several stages between the investigator's original formulation of objectives and the final product that is used for data collection. For example, exploratory interviews and group discussions may be conducted to provide guidance on how to word questions that will communicate accurately what is wanted and that will be acceptable to respondents. The recent developments in questionnaire design laboratories using "think-aloud" and other techniques are an important advance in these developmental stages of questionnaire construction (Lessler, Tourangeau, and Salter 1989; Willis, Royston, and Bercini in press). No matter how extensive the developmental work in questionnaire construction is, however, there remains the need to test the resulting questionnaire under field conditions before it is finally adopted for the actual survey. It is this pretest stage that is the subject of the research reported here.

The purposes of the pretest are to evaluate the individual questions and to ascertain whether they form a cohesive, smoothly flowing questionnaire. Various types of problems may arise with individual questions, among which the following are most frequent:

- Interviewers may have difficulty asking the question because of complex sentence structure, or words that are difficult to pronounce, etc;

[1] Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A.

Acknowledgement: This project was supported by Grant No. H505616 from the National Center for Health Services Research and Health Care Technology Assessment. Other members of the research team for the project were Floyd J. Fowler of the Center for Survey Research at the University of Massachusetts and Katherine Bischoping of the Survey Research Center at the University of Michigan.

- Respondents may have difficulty comprehending the question because the vocabulary is too difficult, the sentence structure is too complex, or because the question does not specify clearly what information is wanted or the form in which it is to be reported;
- There may be a lack of a common understanding of terms and concepts because they are understood differently by different respondents, or are not interpreted as the researcher intended;
- Respondents may have difficulty in cognitive processing of information because they are unable or unwilling to exert the level of effort needed to provide an adequate answer, or because the information is simply inaccessible to them.

Pretests typically involve completion of around 25 to 75 interviews by a few experienced interviewers. The questions are usually evaluated based on interviewers' reports given at a debriefing session. Investigators commonly give interviewers the objectives of the questions and instruct them to be alert to problems that respondents appear to have in answering the questions or that they themselves have in asking them. In some pretests interviewers are told to experiment with changed question wordings if there are problems. At the debriefing session a moderator takes the interviewers through the questionnaire question-by-question, with interviewers noting any problems they encountered. Typically, discussion is in terms of whether the questions "worked" or "didn't work." The "didn't work" covers a wide variety of factors including problems with questionnaire layout, interviewer difficulty with reading questions completely and accurately, and respondent problems with question comprehension and the response task. The information given in debriefings is often subjective and unsystematic, hindering the investigator's attempt to make confident judgments about question problems. Debriefings frequently show a lack of agreement among interviewers about problems, and evaluations often are based on one dramatic interview. Fuller discussions of current pretesting procedures are provided by Converse and Presser (1986), DeMaio (1983), and Hoinville et al., (1978).

Although the pretest plays a critical role in identifying question problems, little research has focused on pretesting methods. There is much evidence that problem questions appear in final questionnaires, indicating that the current methods of pretesting are not adequate in identifying and diagnosing problems with questions. The purpose of the research reported in this paper is to develop improved, more systematic, strategies for pretesting for use in regular survey pretests with little or no increase in cost.

This paper describes the use of behavior coding and special probes for improving the effectiveness of pretesting[2]. With behavior coding, aspects of interviewer and respondent behavior in the question-answer process are coded as a means of identifying questions that need to be reworded or redesigned. The special probes are added to the pretest questionnaire to assess respondents' understandings of questions and specific terms in questions, and to investigate response difficulties. Neither of these techniques is new, each of them having been used or suggested for evaluating questions or for other purposes. Neither of them alone will identify all types of problems with survey questions. The aim of the current research is

[2] The research reported here was part of a larger project (Cannell et al. 1989). Other experimental techniques examined included special training to sensitize interviewers to respondent problems and the use of interviewers' ratings of question difficulty.

to develop and integrate the techniques so that they can be used with regular pretesting to identify questionnaire problems in a systematic manner.

The paper is organized as follows: Section 2 provides some general details of the design of the study; Section 3 describes the behavior coding procedures and presents the behavior coding results; and Section 4 describes the use of special probes and their results. Section 5 briefly summarizes our experience with revising questions to deal with the problems identified by these techniques. The final section of the paper presents some concluding remarks and suggestions for incorporating these experimental procedures in regular pretesting.

2. Study Design

For this study a questionnaire that spanned the range of topics commonly covered in health surveys was constructed. It contained 60 questions about medical and dental care; health care plan membership; health status; nutrition; exercise; knowledge concerning cancer and heart disease risks, and the transmission of AIDS. The questions were taken from questionnaires used in major health surveys. All had been subjected to usual pretesting procedures.

The questionnaire was administered by telephone to 164 respondents. While pretests ordinarily do not use representative samples (because of the small numbers involved and because population estimates are not a pretest goal), they should include respondents from the major segments of the population that are to be sampled in the full survey. In this study, a probability sample of telephone numbers was drawn from lists of telephone subscribers in southeastern Michigan. Respondents at these telephone numbers were selected by a non-random procedure that was designed to yield a balance between male and female respondents and between different age groups. Up to five attempts were made to contact each sampled household to identify a potential respondent. However, minimal attempts were made to persuade reluctant individuals to participate, and if the interview could not be completed at the first contact, only one further contact was attempted. All interviews were tape recorded (with the permission of the respondents).

Sixty of the respondents were interviewed using standard pretest procedures and regular interviewing techniques. Six telephone interviewers with varying lengths of experience each took ten interviews. Interviewer and respondent behaviors in these interviews were coded from the tape recordings.

The remaining 104 respondents were interviewed with the same basic questionnaire. However, for this group special diagnostic probes that aimed to identify problems with questions were added. Some unobtrusive probes were included during the interview; others that had a more intensive focus on special issues were added at the end of the regular interview. Nine different interviewers were used.

The behavior coding yielded indicators of problems with questions. These indicators, along with special probe results, were used to identify a number of questions with significant problems. On the basis of what we learned we revised the questionnaire and took 100 more interviews, using both behavior coding and special probes.

3. Analysis of Interview Behavior

3.1. Behavior coding

Behavior coding was first used in surveys to monitor and evaluate interviewer performance (Cannell, Lawson, and Hausser 1975) and subsequently to investigate the question-answer process more generally (Cannell and Robison 1971; Marquis 1969, 1971a, 1971b; Morton-Williams 1979; Morton-Williams and Sykes 1984; Mathiowetz and Cannell 1980; Dijkstra, Van der Veen, and Van der Zouwen 1985; Groves, Kalton, Oksenberg, and Welch forthcoming). Behavior coding has been used previously by Sykes and Morton-Williams (1987) for evaluating questions for pretests. The coding scheme employed for the current research was developed and adapted from the coding schemes used in earlier research, with the aims of identifying questions that are problems for interviewers or respondents and of diagnosing the nature and source of the problems. The goal was to devise a scheme that could be easily applied in regular pretests. For this purpose, the coding needed to be able to keep pace with the interviewing, and it needed to permit easy aggregation of results for all interviews.

From the beginning it was apparent that coding all interviewer and respondent behavior was too time-consuming and was also unnecessary. Since after the initial asking of the question interviewer behavior tends to be reactive to respondent behavior (e.g., if the respondent gives an inadequate answer, the interviewer probes), it was found that coding interviewers' subsequent behaviors was superfluous. The coding scheme was therefore confined to codes relating to the accuracy and completeness with which interviewers asked the question initially and various respondents' behaviors in answering the question, both initially and after feedback from the interviewers.

Figure 1 gives a brief description of the codes. It should be noted that since accuracy or completeness of answers could not be assessed, the adequate answer code merely means that the answer appeared to meet the question objective. Since interviewers and respondents take turns speaking, respondent behaviors were coded turn-by-turn. Respondent behavior within a turn could involve multiple codes, in which case all the codes were recorded.

Interviewer question-reading codes

E	Exact	Interviewer reads the question exactly as printed.
S	Slight change*	Interviewer reads the question changing a minor word that does not alter question meaning.
M	Major change*	Interviewer changes the question such that the meaning is altered. Interviewer does not complete reading the question.

Respondent behavior codes

1	Interruption with answer*	Respondent interrupts initial question-reading with answer.
2	Clarification*	Respondent asks for repeat or clarification of question, or makes statement indicating uncertainty about question meaning.
3	Adequate answer	Respondent gives answer that meets question objective.
4	Qualified answer*	Respondent gives answer that meets question objective, but is qualified to indicate uncertainty about accuracy.
5	Inadequate answer*	Respondent gives answer that does not meet question objective.
6	Don't know*	Respondent gives a "don't know" or equivalent answer.
7	Refusal to answer*	Respondent refuses to answer the question.

* Indicates a potential problem with the question.

Fig. 1. Behavior code categories

Three experienced telephone interviewers (not those used as interviewers in the study) were employed as coders, each coding approximately equal numbers of interviews from each interviewer. Since the coders were well familiar with interviewing techniques, their training as coders was efficiently accomplished in a few hours.

Many different indicators of problems

with questions can be constructed from the behavior coding. A number of them were investigated and compared before the final set of indicators was chosen. One approach was to analyze only the first respondent behavior coded on the presumption that the first reaction to a question was the most likely to reveal problems. This approach was discarded because it was found that respondents sometimes gave an adequate answer followed by a problem indication. Another approach was to count the number of times each code was used for a question. This approach was discarded because multiple uses of codes were infrequent, and the multiple uses of a code added little additional information to whether or not the code was used at all. The approach finally adopted for the problem indicators was simply to establish whether or not a code was assigned at all for a particular question.

Eight problem indicators were computed for each question corresponding to the eight problem codes given in Figure 1 (i.e., excluding the non-problem codes of exact question reading by the interviewer and adequate answer by the respondent). The reliability of these indicators was assessed from a small-scale coding reliability study conducted for a sample drawn from all the interviews in the research project, including both the interviews with the original and the revised questionnaires. A sample of 19 interviews, involving 1,098 question askings, was independently coded by one of the coders and by a member of the study staff who had been centrally involved in developing the coding scheme and training the coders. The inter-coder reliability for a particular problem indicator was measured by the kappa statistic

$$\kappa = \frac{P - p_e}{1 - p_e}$$

where P is the proportion of all question askings for which the coder and staff member agreed on the presence or absence of the indicator, and p_e is an estimate of the expected proportion of chance agreements. Fleiss (1981) states that values of kappa above 0.75 represent excellent, and values from 0.40 to 0.75 represent fair to good, agreement beyond chance. The values of kappa for the problem indictors were as follows: Slight change, $\kappa = 0.73$; Major change, $\kappa = 0.72$; Interruption with answer, $\kappa = 0.90$; Clarification, $\kappa = 0.93$; Qualified answer, $\kappa = 0.56$; Inadequate answer, $\kappa = 0.85$; and Don't know, $\kappa = 0.86$. The kappa statistic was not calculated for refusal to answer because this code was rarely assigned in the study. As can be seen from the kappa statistics, all the problem indicators, apart from qualified answer, had very good or excellent reliabilities.

3.2. *Application of behavior coding*

Table 1 displays the percentage of the 60 behavior-coded interviews in which each of the problem indicators was assigned. The table shows that many of the questions had sizeable levels of slight changes in question wording. Using an arbitrary index that if 15 percent or more of the respondents had a problem with a question then this indicates a "high level" of the problem, 18 of the 60 questions had high levels of slight changes in question wording. In contrast, major wording changes (including incomplete readings) were relatively rare. As the table shows, a quarter of the questions had high levels of respondent requests for clarification or repeat of the question. The most prevalent problem indicator was inadequate answers, with over two-thirds of the questions having high levels on this indicator. Qualified answers were less common. "Don't know" answers were relatively uncommon and refusals to answer were practically nonexistent.

Table 1. Mean levels and distributions of problem indicators for the 60 questions in the original questionnaire

Problem indicator	Mean level over the 60 questions[1]	Distribution of problem indicators[2]						
		0–4%	5–9%	10–14%	15–19%	20–24%	25–34%	35% +
Interviewer question-reading behavior:								
S Slight changes	12%	18	12	12	5	3	7	3
M Major changes	4%	41	13	3	3	0	0	0
Respondent behavior:								
1. Interruption	4%	47	6	2	2	1	1	1
2. Clarification	10%	20	10	15	9	0	6	0
4. Qualified answer	7%	37	11	3	3	1	3	2
5. Inadequate answer	24%	4	13	7	10	5	6	15
6. "Don't know"	4%	45	8	4	1	0	1	1
7. Refusal	0%	60	0	0	0	0	0	0

Note: The table is based on 60 interviews.
[1] Entries are the percent of times the problem indicator was assigned over all 60 questions.
[2] Entries are the number of questions (out of 60) with problem indicator scores in the specified ranges of percentages.

Table 2 illustrates the variety of patterns of problem indicators for a selection of individual questions. Only the first question example appears to be almost problem free. This was an open question about the respondent's last doctor visit: "What was the purpose of that visit?" Each of the other questions is subject to a sizeable percentage

Table 2. Problem indicator levels for a selection of questions

Problem indicator	Question example number							
	1	2	3	4	5	6	7	8
Interviewer question-reading behavior:								
S Slight changes	2	8	30	8	3	7	5	0
M Major changes	3	19	17	0	2	0	8	2
Respondent behavior:								
1. Interruption	0	35	23	0	0	0	0	0
2. Clarification	2	3	10	3	3	27	30	10
4. Qualified answer	5	3	3	22	27	12	3	0
5. Inadequate answer	5	8	17	13	87	18	30	77
6. "Don't know"	0	0	8	3	12	40	3	5
7. Refusal	0	0	2	0	0	0	0	0

Note: The table is based on 60 interviews. Questions are identified by the example numbers in the text.

of problems. Examples 2 and 3 have high levels of major wording changes and of respondent interruptions, whereas other questions had few such problems but had high levels on other problem indicators such as qualified answers, inadequate answers, and requests for clarification.

Over one-third of the respondents interrupted the interviewer's reading of question example 2 and almost a quarter interrupted with example 3. The two questions read:

> Example 2. "How long ago was the last time you were actually seen by a doctor about your health – within the last month, 1 to 6 months ago, 6 months to a year ago, or more than a year ago?"
>
> Example 3. "How much did you pay, or will you have to pay, *out of pocket* for your most recent visit? Do not include what insurance has paid for or will pay for. If you don't know the exact amount, please give me your best estimate."

In both cases, respondents probably often think that the question has been completed before the interviewer has finished reading it, or they feel that they have heard enough to answer it.

Example 4, which received a high level of qualified answers, read:

> Example 4. "Would you say you are physically more active, less active, or about as active as other persons your age?"

This question poses the challenging task of integrating and evaluating considerable amounts of information to make the comparison. Many respondents qualified their answers to express uncertainty.

In the above examples, examination of the problem indicators and the questions themselves was sufficient to identify the probable sources of the problems. In other cases, and especially when the problem indicators with high levels were requests for clarification, inadequate answers, or "don't know" answers, this procedure was often less effective. Additional, more specific, information was needed to identify the sources of the problems in such cases. Consultation with coders and staff members familiar with the interviews, and the answers to the special probes, proved useful in generating hypotheses about the sources of the problems where they were not apparent from the problem indicators and the question itself.

Example 5 is a question with high levels of both qualified and inadequate answers:

> Example 5. "When was the last time you had a general physical examination or checkup?"

Three problem sources were hypothesized for this question. One, revealed by the answers to a special probe question, is that the concept "general physical examination" lacks clarity. The other two, suggested by those familiar with the interviews, are that the questions lacked a specified response form (e.g., in terms of the calendar month and year, the number of months or years since the examination, etc.) and that respondents had difficulty recalling precisely when the event occurred.

Behavior coding revealed considerable difficulty with example 6 which read:

> Example 6. "What do you think are the warning signs or symptoms of cancer?"

As Table 2 shows, about one in four respondents requested clarification, one in five gave inadequate answers, and two in five gave "don't know" responses. We hypothesized that these problems stemmed partly from a lack of clarity of "warning signs or symptoms of cancer" and partly from the demands on respondents' knowledge and recall. The lack of clarity may be overcome by revised question wording but, since the goal of the question is to assess respondents' knowledge, nothing can be done to simplify the reporting task in this case.

The question of example 7 read:

Example 7. "About how long has it been since you were last treated or examined?"

The question referred to dental care, which was the subject of the preceding question. About a third of the respondents requested clarification and about the same proportion gave inadequate answers to this question. A likely reason for the requests for clarification is that respondents had lost track of the previous question, and did not know what kinds of treatments or examinations were meant. Additionally, like example 5, the question did not specify the form the response was to take, which could account for the high levels of inadequate answers. A third potential source of difficulty was the question sequencing. The preceding question asked for the number of visits for dental care in the past year. While there is no logical problem with next asking when the last visit was, the sequencing appeared to puzzle respondents, especially those who had reported visits to the preceding question.

The last example in Table 2 is from a set of questions sharing the same response categories:

Example 8. "I am going to read a list of things which may or may not affect a person's chances of getting *heart disease.* After I read each one, tell me if you think it definitely increases, probably increases, probably does not, or definitely does not increase a person's chances of getting heart disease. First...

a. cigarette smoking?
b. high blood pressure?
c. diabetes?
d. being *very* overweight?
e. drinking coffee with caffeine?
f. eating a diet high in animal fat?
g. high colesterol?"

Even though interviewers could repeat the response categories, the questions in the set had high levels of inadequate answers, all above 50 percent. (Results in the table are for "high blood pressure?") The response categories appeared to be poorly designed. Answers were inadequate primarily because respondents merely said "definitely" or "probably," which did not serve to single out one of the response choices.

4. Analysis of Special Probes

4.1. Strategies for using special probes

Several researchers (e.g., Cantril and Fried 1944; Schuman 1966; Belson 1981; and Smith 1989) have used special probe questions to explore respondents' understandings of questions. Special probes were employed in this study because they have the potential to supplement the information from behavior coding by providing indications of the sources of problems, and because they may reveal problems not evident in the response behavior. In particular, behavior coding will not detect problems when respondents give acceptable answers to questions that they have misinterpreted or when they choose to give answers to, rather than seek clarification about, questions that they have failed to understand.

The major drawback to using special probes is that only a few questions can be probed without unduly lengthening the interview, and only a few questions can be probed immediately following responses without the danger of influencing responses to subsequent questions. Special probes added at the end of the questionnaire do not have the latter disadvantage, although here problems of retrospection occur.

In order to employ special probes with a large number of questions without unduly lengthening the interviews, three forms of the original questionnaire and of the revised questionnaire were prepared, with different sets of questions being probed on each form. In this way, just over a third of the questions in each form were probed, with approximately 33 respondents receiving a particular probe.

Each questionnaire form included some special probes to be asked immediately following particular questions. Such probes were included for four or five questions and were designed to fit into the flow of the interview without influencing responses to subsequent questions. Some of these probes resembled those routinely used by interviewers (e.g., "Could you tell me more about that?").

Additional special probe questions followed the main body of the questionnaire. Included here was intensive probing of single questions as well as probes that might have disturbed the interview if used earlier. The interviewer introduced this portion of the interview with a version of the following statement:

> "The questions we've been asking you are important for finding out about people's health. We want to make these questions as clear and easy to answer as possible. We would like your help in making them better. To do this, I'd like to read some of the questions I asked you earlier and get some of your thoughts about them."

The purpose of the introduction was to encourage respondents to assume a new role: to become informants rather than respondents. In the informant role, respondents were asked to talk about their interpretations of the question and to report their experiences and difficulties in responding.

Probes were directed at three kinds of problems: comprehension of the question, information retrieval, and (for closed questions) response category selection. In addition, some more general probes – that is, probes not targeted to any particular type of problem – were used. Each form of the questionnaire included a variety of probes, used with a variety of question types.

4.2. Results for the special probes

This section describes the use of a range of special probe questions to identify and diagnose problems with survey questions. Some of the probes used in the original questionnaire appeared to have been ineffective. These probes were replaced in the revised questionnaire by others that seemed to have greater promise. The examples given below are taken from both questionnaires. They are selected to represent the various types of probes that were employed: comprehension probes, information retrieval probes, response category selection probes, and general probes.

4.2.1. Comprehension probes

Comprehension problems may arise because respondents find a question confusing and realize that they do not understand it adequately, or they may feel confident that they understand a question but in fact misinterpret it. Comprehension problems were probed in three main ways: one type of probe asked for respondents' interpretations of the meanings of a particular concept in a question; a second type asked respondents to elaborate on particular aspects of their answers; and a third type asked respondents how clear a particular concept was to them, or about any difficulty they had in understanding the concept.

Comprehension probes revealed substantial misinterpretation and misunderstanding of questions. Fifteen percent or more of the respondents were found to have problems with 12 out of the 18 questions probed.

The probes revealed misinterpretations of key terms in the questions, but did not reveal uncertainty or confusion about question meaning. Respondents did not appear to doubt their own, often mistaken, interpretations.

The following question provides a striking example of the success of a probe asking for a conceptual interpretation. The question read:

"During the past 12 months, that is, since January 1, 1987, *about* how many days did illness or injury keep you in bed more than half of the day?"

This question was probed at the end of the interview. One probe was "How clear was it to you what to include as a *half a day in bed*?" Most of the respondents who volunteered a definition interpreted this to mean not getting out of bed in the morning and staying in bed until noon or later. Others gave lengths of time, from 2–4 hours up to 12 or more hours. Another probe for the same question was: "What if you were staying in bed because you felt you were coming down with something. Would you count that as staying in bed because of illness?" About two-thirds of the respondents would include this as illness while the others would not. The differing interpretations revealed by responses to these and other similar probes indicate considerable variability in interpretations of question meaning.

Another question that responses to special probes showed was not understood in the same way by different respondents was:

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

One probe to this question asked respondents to identify which health professionals they would include as doctors or assistants from a list that included chiropractors, physical therapists, podiatrists, optometrists, psychiatrists, nurses, and laboratory or x-ray technicians. There was considerable disagreement among respondents for each of these health professionals as to whether they should be included as "doctors or assistants." Responses to another special probe revealed that about a third of the respondents thought that medical advice obtained on the telephone should be included as instances of having "seen or talked to a doctor or assistant about your health", whereas the remainder disagreed.

The next question demonstrates the effectiveness of comprehension probes asking respondents to elaborate on particular aspects of their answers:

"In the past 4 weeks, Monday (DATE 4 WEEKS AGO) and ending this past Sunday (DATE LAST SUNDAY), have you done any exercise, sports, or physically active hobbies?"

Respondents who answered "no" to that question were asked at the end of the interview:

Probe: "...You said that in the past 4 weeks you had not done any exercise, sports, or physically active hobbies. Did you get any exercise at all during that time?"

About a third of those who initially reported no exercise nonetheless mentioned exercise (primarily walking) in response to the special probe. While these respondents appeared not to consider walking as real exercise, others did. About a third of those who initially reported exercise mentioned walking in response to the special probe, "You said that in the past 4 weeks you had done some exercise, sports, or physically active hobbies. Could you tell me more about that?"

Another question that was probed in a similar way was:

"When was the last time you had a general physical examination or checkup?"
Probe: "What was the main reason you went for that visit?"

Responses to the probe indicated that many respondents reported visits to "check up" on a particular health condition or for a specific test or examination. According to question objectives, these should not have been included.

Comprehension probes asking about difficulties or trouble in understanding questions revealed fewer problems than other comprehension probes. The reason for this is unclear. Perhaps the probes soliciting reports of trouble or difficulty happened to be used with questions without such problems. Or, perhaps respondents are reluctant to admit to problems, seeing it as reflecting poorly on their abilities. Another reasonable explanation is that respondents' definitions of problems or difficulty are different from researchers' definitions. Respondents may not consider themselves as having difficulty understanding questions, even when they request clarification. However, when probed to find out how they understood questions, they reveal misunderstandings and lack of agreement about question meanings.

It is instructive to compare the effectiveness of the behavior coding and the comprehension probes in identifying comprehension difficulties with the 18 questions with which comprehension probes were used. Fifteen percent or more of respondents were found by the probes to have comprehension difficulties for 12 of these 18 questions. For 5 of the 12 questions, the behavior coding showed that 15 percent or more of respondents asked for clarification, but the other 7 questions were not detected as having clarification problems. None of the questions classified as nonproblematic according to the comprehension probe was classified as problematic according to the behavior coding. This finding is consistent with respondents being largely confident (but often incorrect) as to question meaning, or being reluctant to admit their uncertainty about it. In this situation comprehension probes may serve to reveal comprehension difficulties that are missed by behavior coding.

4.2.2. Information retrieval probes

Fifteen questions were probed for difficulties with recalling or organizing information. Some of the probes asked respondents to talk about how they arrived at their answers, or to report problems they had in answering. For example, the information retrieval probe used with a question asking how long it had been since the respondent had last been treated or examined for dental care was "How did you figure out when that was?" Others asked respondents how hard it was for them to answer, or asked them to assess the accuracy of their answers (for example, "Do you think your answer was exact, pretty close, or not very close to the actual time?").

Of the 15 questions with which information retrieval probes were used, only one had more than 15 percent of respondents reporting difficulties. One possible explanation for the paucity of evidence of retrieval problems is that the questions actually caused few problems for respondents. The coding of respondent behavior in the interviews, however, revealed that ten of the 15 questions had high levels of behaviors often associated with retrieval problems – inadequate, qualified, or "don't know" answers. A more likely explanation is that respondents generally do not see themselves as having problems in giving answers, even when their interview behavior suggests otherwise. For example, when a respondent gives an inadequate answer, this is no problem for him or her. From the researcher's viewpoint, however, inadequate answers indicate a problem with the question. It also is possible that better probes could be devised, although what they would be is not obvious.

4.2.3. Response category selection probes

While respondents might retrieve the information needed to answer a closed question, they might have difficulty mapping that information into the response choices provided. Response category selection probes were designed to reveal this type of problem. For a question asking how much of the time during the past month the respondent had been a happy person – "all of the time, most of the time, a good bit of the time, some of the time, a little bit of the time, or none of the time" – the probe was:

Probe: "In answering that question, how hard was it for you to pick an answer that describes how you really felt?"

Six closed questions were probed for respondent difficulties with selecting the appropriate response category. Although responses to these probes gave evidence of other difficulties, they failed to reveal the particular type of problems for which the probes were designed. The reasons for this failure are unclear. It may be that respondents did not have response mapping problems, or it may be that they did not understand the probes as we intended. Upon reflection, we think it is difficult to phrase probes for this type of problem without giving extended explanations.

4.2.4. General probes

General probes were employed with 12 questions with the aim of acting as a general stimulus for additional information. These probes were variations on "Could you tell me more about that?"

The general probes indicated significant levels of comprehension problems with two of the 12 questions. One question asked respondents which of two statements they agreed with most: "(A) What people eat or drink has little effect on whether they will develop major diseases; or (B) By eating certain kinds of foods, people can reduce their chances of developing major diseases." Responses to the general probe indicated that many respondents interpreted the second statement to include avoidance of certain foods. The other question asked respondents to rate their health on a three-point scale, compared to others their age. In this case the responses to the general probe appeared to show that a number of respondents rated their health in some absolute sense, rather than compared to others their age.

The behavior coding also identified these two questions as problematic. However, for one of them the behavior coding showed high levels of qualified answers – a type of answer more likely to reflect retrieval problems than comprehension problems as revealed by the general probe. For four other questions, behavior coding results revealed some sort of problem, whereas the general probes revealed none.

It is difficult to draw conclusions from the small number of problems identified by the general probes. It may be that these probes are often too non-specific and are not sufficiently directed toward potential problem sources. Sometimes, however, the original question provides an adequate frame of reference for a general probe, so that it yields useful information. This probably was the case with the two questions for which general probes revealed problems. Based on responses to the probes for these questions, it also appears that general probes may be more useful for revealing comprehension problems than other problems.

5. Revision of Questions

Following the analysis of the first interviewing phase, we revised the questions identified by the behavior coding and special probes as having problems, and tested the revisions. We were able to make considerable progress in reducing scores on the problem indicators. We were most successful in improving the questions to reduce interruptions, qualified answers and clarifications to tolerably low levels.

Interruptions were reduced by rearranging components of the question so that it did not appear to be completed prematurely. For example, revising example 2 to, "Was the last time you actually saw a medical doctor about your health within the last month, 1 to 6 months ago, 6 months to a year ago, or more than a year ago?", reduced interruptions substantially. The same was true for the revision of example 3: "The next question is about how much it cost you or your family for your most recent visit to a medical doctor. Not including what insurance pays, about how much did you pay or will you pay for the visit?"

Where exact answers were not required, levels of qualified answers were reduced by simplifying the reporting task to allow estimates as well as exact answers. When example 4 was revised to, "Thinking about physical activity, would you say you probably are *more* active, *less* active, or *about as* active as other persons your age?", hardly any respondents indicated uncertainty about the answer.

Requests for clarification were reduced by rewordings to provide clearer descriptions of the concepts. While example 7 followed a question about dental care, example 7 itself did not explicitly state that it too referred to dental care. The revision was intended to clarify this: "Was the last time you were treated or examined for dental care within the last 2 weeks, more than 2 weeks to 6 months ago, or more than 6 months ago?" (Respondents who had reported no visits in the last year to the earlier question instead were asked, "About how many years ago was the last time you were treated or examined for dental care?") Levels of requests for clarification were considerably reduced for these questions. Levels of requests for clarification were reduced somewhat with the revision of example 6: "Now we want to get some of your ideas about symptoms of cancer. What are some of the symptoms that a person should be concerned about because they may be warning signs of some kind of cancer?"

We generally were less successful in reducing levels of inadequate and "don't know" answers. While there were sizeable reductions in levels of inadequate answers, they did not achieve tolerably low levels. This was the case with the revision of example 7 (above) as well as with the revision of example 8, which used the response choices, "large effect, some effect, little effect, or no effect." We were quite unsuccessful in reducing levels of "don't know" answers. This may reflect enduring difficulties with recalling and organizing information as required by the reporting tasks. Generally speaking, it is difficult to simplify the reporting task without at the same time modifying the question objective.

Some questions resist significant improvement. These include questions containing complex or fuzzy concepts that defy simplification or clarification, and questions involving very difficult reporting tasks, placing unacceptable demands on respondents' knowledge, recall, ability, or organizing capacities. For such questions some improvement may be achieved by rewording, but no amount of revision can solve the underlying problems. The most feasible solutions to these problems are either to revise the statement of data required, or to frame multiple questions in place of the single question. Thus, for example, in this study "HMO" was found to be a fuzzy concept. It would take a battery of questions to identify HMO visits, and even then the respondent may not be able to provide the answers.

6. Conclusion

Although the questionnaire is the measuring instrument upon which the success of the whole survey ultimately depends, its development and testing are the least scientifically rigorous components of the survey process. Regular pretesting, with a reliance on interviewer debriefing to detect problems with questions, is an unsystematic procedure that fails to uncover many problems (Bischoping 1989). The objective of the behavior codings and special probes employed in this study is to provide more systematic and objective information for evaluating questions.

Our study revealed many problems with questions that are widely used and that have been subjected to regular pretesting. This raises the question of how significant are the problems that the behavior coding and special probes have identified. Since the ultimate concern is whether respondents comprehend and answer questions as the researcher intended, the basic issue is whether the problems affect the answers respondents give. We found a number of cases where the problems did appreciably affect respondents' answers. In particular, the behavior coding and special probes identified several cases of unclear concepts. When these concepts were clarified in the revised versions of the questions, the distributions of answers obtained were often markedly different (Fowler 1989a). For example, revising a question about exercising or playing sports regularly to explicitly include walking and to clarify the meaning of "regularly" substantially increased the proportion of positive responses, whereas revising a question about butter consumption to explicitly exclude margarine led to a sizeable decrease in the reports of butter consumption.

A detailed behavior coding procedure was employed for this experimental study in order to determine which aspects of interview behavior are most effective in identifying question problems. Based on the analyses of the detailed codes, it is clear that a much simpler coding scheme can serve well. The behavior coding that we are currently using as a standard pretest procedure has evolved into a simple, low-cost, flexible system, with the possibility of adapting the codes to address special issues with certain questions. Codings of telephone interviews can be performed either live or from tape recordings. When tape recordings are used, it is rarely necessary to stop the tape in performing the coding. Coders enter the codes directly into a personal computer, and a simple program enables the distributions of codes to be produced shortly after the pretest interviewing is completed. In this way, the coding results are available to provide a basis for the interviewer debriefing.

One limitation of behavior coding is that it does not always identify the sources of the problems it uncovers. By using the coding results as the basis for the interviewer debriefing, discussion can be directed toward identifying the problem sources. An alternative method for identifying problem sources is to hold a debriefing session with the coders. We have found that coders have a more comprehensive view of the interview and are especially attuned to the problems identified by the coding. They can also provide objective analyses of questions without personal involvement. Our current practice is to debrief both the coders and the interviewers about their coding and interviewing experiences, respectively, using the behavior coding results as background information for directing the discussions.

Another limitation of behavior coding is that it does not uncover all problems. The particular strength of special probes lies in their ability to reveal problems that are not evident in interview behavior. Our experience with special probes was mixed. They worked well for comprehension problems where we had an idea of what concepts might be troublesome. In such cases, they are a valuable addition to pretesting techniques. The special probes we employed to try to identify other types of problems were less successful. It may be that more effective probes can be devised to reveal these other types of problems. We are continuing to explore this possibility.

Other techniques included in this study, but not reported here, are special training for pretest interviewers in how to recognize problems with questions and the use of rating forms on which the pretest interviewers rate the questions for problems on the basis of their pretest experiences. These techniques, discussed by Fowler (1989b), also offer some significant potential for improving pretesting. As Fowler notes, regular survey interviewers are trained in skills of asking questions, probing, etc., but different skills and sensitivities are required for interviewers to recognize their own difficulties with asking questions or identify respondents' problems. If interviewers could identify problem questions reliably, their ratings could provide a useful supplement to the behavior coding.

We conclude from this study and our more recent experience that the addition of these techniques makes a significant improvement to standard pretesting. As experience is gathered in the use of behavior coding, special probes, and interviewer ratings, these techniques should prove even more valuable in the future.

7. References

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

Bischoping, K. (1989). An Evaluation of Interviewer Debriefing in Survey Pretests. In C. Cannell et al., (eds.), New Techniques for Pretesting Survey Questions, Chapter 2. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Cannell, C.F. and Robison, S. (1971). Analysis of Individual Questions. In J.B. Lansing, et al. (eds.), Working Papers on Survey Research in Poverty Areas, Chapter 11. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). A Technique for Evaluating Interviewer Performance. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F.J. (1989). New Techniques for Pretesting Survey Questions. Research Report. Survey Research Center, The University of Michigan.

Cantril, H. and Fried, E. (1944). The Meaning of Questions. In H. Cantril (ed.), Gauging Public Opinion. Princeton: Princeton University Press.

Converse, J. and Presser, S. (1986). Survey Questions: Handcrafting the Standardized Questionnaire. Quantitative Applications in the Social Sciences No. 63. Beverly Hills, CA: Sage Publications.

DeMaio, T. (ed.) (1983). Approaches to Developing Questionnaires. Statistical Policy Working Paper 10, Office of Information and Regulatory Affairs, Office of Management and Budget.

Dijkstra, W., Van der Veen, L., and Van der Zouwen, J. (1985). A Field Experiment on Interviewer-Respondent Interaction. In Brenner et al. (eds.), The Research Interview, Chapter 3. London: Academic Press.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions. 2nd edition. New York: Wiley.

Fowler, F.J. (1989a). Coding Behavior in Pretests to Identify Unclear Questions. Proceedings of the Fifth Conference on Health Survey Research Methods, F.J. Fowler (ed.), 9–12. Washington D.C.: National Center for Health Services Research and Health Care Technology Assessment.

Fowler, F.J. (1989b). Evaluation of Special Training and Debriefing Procedures for Pretest Interviews. In C. Cannell et al., (eds.), New Techniques for Pretesting Survey Questions, Chapter 4. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Groves, R., Kalton, G., Oksenberg, L., and Welch, D. (forthcoming). Linked Telephone Surveys: A Test of Methodology. Vital and Health Statistics, Series 2. Washington, D.C.: National Center for Health Statistics.

Hoinville, G., Jowell, R. and associates (1978). Survey Research Practice. London: Heinemann Educational Books.

Lessler, J., Tourangeau, R., and Salter, W. (1989). Questionnaire Design in the Cognitive Research Laboratory. Vital and Health Statistics, Series 6, No. 1. Washington, D.C.: National Center for Health Statistics.

Marquis, K.H. (1969). Interviewer-Respondent Interaction in a Household Interview. Proceedings of the Social Statistics Section, American Statistical Association, 24–30.

Marquis, K.H. (1971a). Purpose and Procedure of the Tape Recording Analysis. In J.B. Lansing, et al. (eds.), Working Papers on Survey Research in Poverty Areas, Chapter 10. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Marquis, K.H. (1971b). Effects of Race, Residence and Selection of Respondent on the Conduct of the Interview. In J.B. Lansing, et al. (eds.), Working Papers on Survey Research in Poverty Areas, Chapter 12. Ann Arbor, MI: Survey Research Center, The University of Michigan.

Mathiowetz, N. and Cannell, C. (1980). Coding Interviewer Behavior as a Method of Evaluating Performance. Proceedings of the Section on Survey Research Methods, American Statistical Association, 525–528.

Morton-Williams, J. (1979). The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire. Quality and Quantity, 13, 59–75.

Morton-Williams, J. and Sykes, W. (1984). The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Questions. Journal of the Market Research Society, 26, 109–127.

Schuman, H. (1966). The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, 31, 218–222.

Smith, T.W. (1989). Random Probes of GSS Questions. International Journal of Public Opinion Research, 1, 305–325.

Sykes, W. and Morton-Williams, J. (1987). Evaluating Survey Questions. Journal of Official Statistics, 3, 191–205.

Willis, G.B., Royston, P., and Bercini, D. (in press). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. Applied Cognitive Psychology.

Received December 1990
Revised June 1991