FINAL REPORT
August, 1989

NEW TECHNIQUES FOR PRETESTING SURVEY QUESTIONS

Charles Cannell Lois Oksenberg
Graham Kalton Katherine Bischoping
Survey Research Center,
The University of Michigan,
Ann Arbor, Michigan

Floyd J. Fowler
Center for Survey Research,
University of Massachusetts,
Boston, Massachusetts

ACKNOWLEDGMENTS

In a research endeavor such as this many people make significant contributions. We wish to recognize especially the contributions of Steven Blixt who supervised and administered the data collection, and conducted the debriefing sessions, Marion Wirick who organized, edited, and handled the word processing of the final report at SRC; and of Pam Sapyta who handled data processing and Anatis Benfield who did most of the typing at CSR. Their help is gratefully acknowledged.

This project was supported by Grant Number HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment.

TABLE OF CONTENTS

CHAPTER 1*

Background and Research Plan

INTRODUCTION

The objective of this research is to develop and test some techniques for pretesting survey questions prior to data collection. While there is a consensus among survey methodologists on the need for question testing, little attention has been given to investigating or even describing adequate procedures. In fact, most of what has been written on the topic consists simply of scattered paragraphs in survey textbooks.

Methods used to develop questionnaires vary among organizations and researchers, depending in part on previous knowledge about surveying the topic, and the experience of the researcher. For illustration, consider steps involved in preparing and testing a questionnaire on a topic with which the investigator has little experience. Although the concepts and objectives are clear for the researcher, developmental work is needed to learn how to operationalize them and how to present them in the questionnaire.

Before questions can be prepared, it is necessary to know the level of respondent knowledge that can be assumed and something of the terminology that respondents will understand. For example, to investigate sources of health re, the researcher would like to know whether the respondent uses an HMO, urgi-care, satellite clinic, outpatient clinic, emergency room, etc. Do respondents differentiate these sources, and can they identify the source of their medical care?

Preliminary exploration to provide a basis for understanding the types and levels of questions that can be asked may involve individual or group interviews and may include studies of information storage and retrieval. Such interviews may be structured or unstructured, face-to-face, or telephone. Knowledge from this operation enables the researcher to specify a set of

*This project was supported by Grant Number HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment. Members of the research team were Charles Cannell, Floyd J. Fowler, Graham Kalton, Lois Oksenberg, and Katherine Bischoping. Dr. Fowler is from the Center for Survey Research, University of Massachusetts. The other members of the research team are from the Survey Research Center, The University of Michigan.

variables needed to measure the concept of interest.[1] It is these variables that designate the information needed and provide the specification for the questions.

Based on the initial explorations, the researcher designs questions that appear likely to obtain the required information. These preliminary questions are then explored in individual or group interviews. The questionnaire development laboratory techniques such as those described by Royston, et al. (1987) are especially useful. Their objective is to ascertain how to refine questions so that they communicate clearly with respondents and elicit responses that fulfill data specifications.

The product of the developmental phase is a set of questions that appear to be understood by respondents and will provide the data needed. The amount of exploration and developmental work needed varies greatly depending on the complexity of the subject matter and previous experience in measuring the variables. For topics with which the researcher has extensive previous experience, little developmental work may be needed. Measures with a history of use and adequacy may be adopted without further development. For topics without such a history, the developmental work also may include major research endeavors. Measures in mental health, use of alcohol or drugs, for example, have taken years of methodological development for valid measurement.

While there is wide variation in the amount and type of developmental activity undertaken, survey researchers agree that the resulting questionnaire must be tested under field conditions before final use. It is this stage of question-testing or "pretesting" that is the subject of the research to be reported here.[2]

The Importance of Question Testing

The interview is a meeting between the world of the researcher and the world of the lay person. The questionnaire serves as a translation of the researcher's goals into the language of the respondent. The ability of the questionnaire to obtain valid and reliable information depends on the quality

[1] For a description, see Converse and Presser, 1986; Hoinville, et al. 1978; and DeMaio, 1983.

[2] Some large surveys designed for collecting descriptive information about a population use additional modes of testing. Surveys such as NHIS, CPS, and the National Crime Survey are conducted primarily to provide distributions of frequency of events and behaviors. They need to test questions but, in addition, need to test procedures for coding and processing the survey data. They also examine the distributions of key variables for some assurance that they appear reasonable. For these pilot studies a few hundred interviews are required to include a number of rare events and to provide some stability of prevalence estimates. These extensions of the usual question-testing are not included in our research.

and appeal of the translation. The respondent must be able to understand the question, retrieve the required information, and be willing to reveal it.

Methodological investigations have demonstrated repeatedly that poorly designed or worded questions are major sources of survey error. It is crucial to the quality and accuracy of a survey, then, that researchers identify problematic questions prior to the major data collection. The pretest is an opportunity to test the adequacy of the questions to achieve their intended goal.

While survey researchers agree on the benefits of pretesting, in practice it is often done poorly or not at all. Hunt (1982) quotes the following comment on pretesting from Backstrom and Hursch (1963):

> "No amount of intellectual exercise can substitute for testing an instrument designed to communicate with ordinary people."

In comments that apply beyond the market research context he refers to, Hunt goes on to say:

> "Despite the generally accepted importance of pretesting, the pretesting process is given short shrift in both marketing research literature and marketing research practice. In practice, the pretesting of questionnaires is always done in a hurried, nonsystematic fashion. As Lehman (1979) has lamented, the pretesting stage in the research process is one 'most likely to be squeezed out to cost/time pressures.' Similarly, reports of research on pretesting are almost nonexistent in the literature of marketing and other social sciences."

Conventional Pretesting

We reviewed four of the most commonly used textbooks of survey methods (Warwick and Lininger, 1975; Moser and Kalton, 1972; Babbie, 1975; Selltiz, et al., 1976). In each case the authors exhort researchers to pretest. However, they give little advice about pretesting procedures and do not explain how questions are to evaluated.

Conventional wisdom on how pretests should be conducted can be summarized as follows.

Sample size. Authors suggest numbers ranging from 10 to 100 for testing questions; the majority suggest 25 to 75 interviews. Some authors suggest that long or complex questionnaires, or questionnaires to be used with unsophisticated populations, need larger pretests. Another suggestion is that investigators with considerable experience in the survey topic may need fewer interviews than less experienced investigators.

Selection of respondents. There is general agreement that the pretest sample should include respondents from the major segments of the population that are to be sampled. Some methodologists argue that the pretest should be weighted toward groups which may have the most difficulty with the questionnaire. This may be on the basis of educational level or other socio-economic variable, or on the basis of particular experiences of importance in the survey; i.e., home owners.

Interviewer selection. There is less consensus on who should conduct the pretest interviews. Some researchers recommend the most experienced and competent interviewers available be involved in pretesting, on the grounds that they will be particularly effective in identifying question problems and suggesting wording changes. Others argue for a mix of experience. The less experienced are more likely to encounter difficulties with situations the more experienced interviewers handle with ease.

Training interviewers for pretesting. The literature contains few specific recommendations regarding the amount and type of interviewer training for pretesting. Generally, it seems to say that standard interviewing techniques should be used and interviewers should be especially sensitive to identifying question problems. Some researchers encourage interviewers to try variations in the question wording to discover better ways of obtaining the information.

Transmitting interviewer experience to decision-makers. There is no generally accepted technique for communicating the pretest experience to the survey director. Both oral and written feedback from interviewing are in use.

Written reports may include copies of the questionnaire with interviewer comments on each question judged to have a problem as well as systematic ratings of question difficulties. Oral feedback may include interviews between individual interviewers and supervisory or research personnel. The most frequently used method is a group discussion with interviewers at the completion of the pretest, with a field supervisor and the survey director leading the discussion of each question to identify and describe problems and their causes.

Our impression of the present state of affairs is that pretests typically involve completion of a small number of interviews by a few experienced interviewers, with the questions evaluated based on interviewers' reports. Interviewers commonly are given the objectives of the questions and instructed to be alert to problems that respondents appear to have in answering the questions or that they themselves have in asking them. At a debriefing session the questionnaire is reviewed question by question with interviewers noting problems. 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 inability to apply their training and read 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 researcher's attempt to make confident judgments about question problems.

PURPOSE AND PLAN OF THIS RESEARCH

As the above description suggests, the critical task of evaluating a measuring instrument is left to the subjective evaluation of the researcher with little objective guidance from the pretest experience. As a result many flawed questions filter into the final survey. The informal, intuitive nature of questionnaire evaluation is unacceptable for researchers devoted to scientific measurement. Our purpose in this study was to develop and test a systematic set of procedures for pretesting survey questions. The goal is to provide survey investigators with techniques that will provide objective measures for identifying question problems, replacing the subjective and vague judgments presently used as standard practice.

Subsequent chapters describe in detail the development and testing of new pretesting procedures. In brief, the techniques explored include the following:

- Coding and analysis of interviewer and respondent behavior to identify problems on a question-by-question basis. Many survey organizations now use some method of observing and categorizing interviewer or respondent behavior that occurs during interviews, the most frequent being to code the interviewer behaviors as a supervisory technique to evaluate interviewing performance. Several studies have included respondent codes to identify question difficulty.[3]

- Use of special probe questions to study respondents' understanding of questions and retrieval of information. We used special probes asking respondents to elaborate their responses or give their interpretations of questions. Such techniques have been used by Schuman (1966). Belson (1981) has made extensive use of such probes to demonstrate respondent misinterpretation of questions.

- Use of interviewers specially trained to recognize problems with questions and to rate questions for these problems. Regular survey interviewers are trained in skills of asking questions, probing, etc., but have minimal experience in awareness of their own difficulties with asking questions or of respondent problems. This requires different skills and sensitivities.

While none of these techniques is new, they have not commonly been used to evaluate survey questions. The objective of this research has been to develop them into an integrated set of procedures that provides better evaluation of questions than do present pretest methods. The procedures should be simple enough to be used on pretests without adding significantly to the time and cost and not require special expertise.

[3] A description of various behavior coding procedures and their uses can be found in Cannell and Oksenberg, "Systematic Observation of Behavior in Telephone Interviews. Chapter in Robert M. Groves, et al., Telephone Survey Methodology, New York, NY, 1988.

The Nature of Question Problems

This section provides an overview of the kinds of problems with questions that are most frequent and indicates how the proposed techniques may help to identify these problems. There are many different types of potential problems in survey questions, and they may be classified in various ways. Both research and experience in conducting surveys have identified four types as probably the most frequent.

Questions that are difficult to ask

Interviewers sometimes have trouble reading questions accurately. Unfamiliar words or words that are difficult to pronounce may cause error, as may phrases that are awkward for the interviewer to read. Perhaps the most common cause of this type of problem is complex sentence structure. The sentence structures used by academics are often awkward for survey questions.

Inaccurate question reading also can occur when interviewers consider that a question is difficult to understand and attempt to help respondents by telling them in different terms what is wanted. A catch-all question like, "Can you tell me the type, coverage, and terms of your current health insurance policies?" invites interviewers to alter the question to make it easier for the respondent to answer its separate components.

We expect that coding of interviewer question-asking will identify questions that are difficult to ask as worded. Interviewers specially trained to attend to reading difficulties are also expected to identify questions with this problem and the source of the problem.

Question comprehension

Comprehension problems occur when a question fails to tell respondents unambiguously what is wanted. Question wording issues that produce such basic ommunication problems are: difficult vocabulary, complex sentence structure, or unclear specification of the required information or the response form.

Coding of interview behavior should identify many difficulties of question comprehension. A respondent who fails to understand a question may ask for an interpretation or give an inadequate answer, and the interviewer may need to engage in probing activity to obtain a codeable response. However, some comprehension problems may not be easily detected by behavior analysis, particularly if they result from difficult vocabulary. For example, most respondents may readily answer "No" to the question "Do you have any allergies?" (especially if the question is preceded by a series of similar questions), even though they do not know what "allergies" means. Health surveys are particularly subject to this type of problem because the general public often does not know illnesses and diseases by the technical terms used by professionals. Behavior coding will uncover such difficulties if a sufficient number of pretest respondents inquire about the meaning of the difficult words, but previous research has shown that respondents seldom reveal their ignorance (Belson, 1981). The special probe technique, used to investigate interpretations of questions and terms, seems likely to be a more investigate interpretations of questions and terms, seems likely to be a more dependable way of identifying terms that are difficult to understand. To the extent that respondents do indicate comprehension difficulties, specially trained interviewers may be able to identify questions causing this problem.

Lack of a Common Understanding of Terms and Concepts

Lack of common understanding occurs when the terms or concepts used in a question have different meanings for different respondents, or when the question is not interpreted as the researcher intended. An example is the word "doctor," which some respondents may interpret to include chiropractors, osteopaths, and naturopaths as the researcher intended, but others may interpret to include only physicians (M.D.s).

The difficulty in identifying this type of problem is that respondents feel comfortable answering the question, assuming that their own understanding of its meaning is the one intended, and therefore gave no indication of a discrepancy. Neither the pretest interviewers nor behavior coding can detect the problem if a meaningful answer is given. Only when a diligent and thoughtful respondent appreciates that a term could be interpreted differently and asks what was intended will the problem be apparent.

We anticipate that the use of special probes for comprehension will prove to be the most effective way of uncovering such problems.

Difficulty in Cognitive Processing of Information

Information processing difficulty is not an issue of comprehension or understanding but of the inability or unwillingness of respondents to respond adequately. Sometimes the information required is not accessible to the respondent and reporting is impossible, i.e., What was your father's income when he was first married? More often, the problem is not that the information is inaccessible, but that considerable effort is needed to retrieve and process it. Considerable research demonstrates that recall of health events becomes difficult as time elapses. Events that have little salience or importance for respondents are also more difficult to recall. The surprising finding is that retrieval becomes difficult much more quickly than might be expected. Reports of doctor visits begin to deteriorate in a couple of weeks (Cannell and Fowler, 1963) and hospitalization (Cannell and Fowler, 1965) reports after two or three months.

Evaluating questions, not for wording but for the ability of the respondent to produce the information requested, is an important goal of pretests. Often respondents will express their difficulty -- "I don't remember," "It was a long time ago," "I can't tell you because it was not very important." Inadequate responses, hesitations in reporting and long pauses before responding may indirectly indicate difficulty. The behavior coding should provide some diagnostic information, and special probes asking about response difficulties may provide additional indicators. Specially trained interviewers also may identify questions with which respondents have this type of problem.

Other Question Problems

Two other types of problems may be identified during pretests, but they are not addressed in this study because they involve the more basic measurement issues of response bias and construct invalidity. Response bias may occur when requested information is perceived by the respondent to be socially undesirable, embarrassing or threatening. The respondent knows what is wanted, has no problems of comprehension or difficulty in processing the information, but is unwilling to report accurately.

The pretest may also make clear that questions do not produce information that adequately measures the intended construct. Questions on job stress may be readily and accurately answered but they may not provide a valid measure of stress as conceived by the researcher, because of an inadequacy in operationalizing the construct. These are not problems of question wording. Solutions require redesign of measures or different techniques of inquiry, rather than simply question rewording.

Study Design

Experimental Groups and Procedures

To develop and test the effectiveness of the three techniques, we conducted a series of pretest interviews. All interviews in the study were taken by telephone and were tape-recorded with respondent permission.* Interviews were completed in two phases, as shown in Figure 1. The "first pretest" included three groups of interviews.

Group 1 was a "standard" pretest; interviewers used standard interviewing techniques, with no attempt during the interview to vary the questions or do extra probing. Interviewers were told to note questions that seemed to cause problems, either for themselves or their respondents. Six interviewers from the Survey Research Center (SRC), The University of Michigan, each took ten interviews for a total of 60 interviews.

The interviewers were selected from the regular SRC interviewing staff. Three had more than six months of interviewing experience and three less than six months. We felt that that number of interviews should approximate a regular pretest. If techniques could not identify problems within that sample size, they would not be useful in usual pretests. These interviews were then behavior-coded and the frequency of behaviors that indicated problems with questions was tabulated.

Group 2 involved use of special probes designed to elicit evidence of respondent problems with questions. Some of these probes were unobtrusive and

*The laws covering recording telephone communications vary from state to state. In Michigan, Massachusetts, and many other states, recording is legal provided the individual gives permission and the permission is included in the recording.

were included during the interview. Others, especially those involving intensive probing of special issues, were used at the end of the interview.

One hundred and four interviews were taken in this group. The interviews were divided into three subgroups, each with a different set of special probes. This was done to avoid burdening the respondent or significantly altering the nature of the interview. The same basic questionnaire from Group 1 was employed. Nine SRC interviewers (five experienced and four less experienced) were used, all different from those in Group 1.

Group 3 used interviewers from the Center for Survey Research (CSR), University of Massachusetts, who were specially trained and given extensive practice to be sensitive to and to identify respondent and interviewer problems with questions. These interviewers rated the questions for various kinds of difficulties upon completion of their interviews. Fifty interviews were taken by five interviewers. These interviews were also behavior coded by SRC coders and the results tabulated to identify question problems.

Following completion of the first pretest interviews from these three groups, the questionnaire was revised based on the question problems that had been identified by the various techniques. Questions with substantial numbers of problems were reworded or redesigned. The revised questionnaire was administered to two additional groups of respondents.

Group 4 included 100 interviews by SRC interviewers using special probing. Probes also were revised on the basis of the first pretest experience. A sample of 60 of these interviews were behavior coded.

Group 5 used specially trained CSR interviewers. None of the second pretest interviewers had participated in the first pretest.

In addition to the special techniques, these groups also involved standard debriefings in which the interviewers met with the study staff and supervisors to discuss the questions. These sessions were tape-recorded.

Sample

The sample for these groups was drawn from a listing of telephone subscribers in southeastern Michigan. Random subsamples were drawn for each group. As is common for pretests, attempts to contact or persuade potential respondents were limited. One follow-up phone call was made if needed and no follow-ups were made of refusals.

Questionnaire

The choice of questions to use for this study required careful consideration and investigation. We needed questions that had actually been used at least at the pretest stage. We could "create" questions, but this would raise issues of research objectivity. We initially attempted to locate a variety of health pretest questionnaires at the pretest phase, but a few inquiries to locate these were not successful.

The decision was to use questions that had been used in various surveys on health topics. Since our goal was to construct a questionnaire that included a variety of typical health survey questions, the first step was to assemble a large pool of questions covering the range of topics commonly included in health surveys. We began by collecting a number of questionnaires that had been used in major health surveys or, in two instances, in pretests for major surveys. Some 15 questionnaires were assembled from such sources as the National Center for Health Statistics, the National Center for Health Services Research, the Canadian Fitness Survey, and several non-governmental organizations such as NORC, Chilton, and Lou Harris. Since most health surveys include one or more of four topics, we decided to select questions from each of those areas:

- Utilization of health services,
- Health insurance,
- Health status,
- Health behaviors and information, including health promotion, disease prevention, and risk factors.

There was considerable overlap among the questionnaires, with questions from one survey adapted or used verbatim in another. This reduced the number of eligible items. Since our goal was to include only questions relevant to all or nearly all respondents, we also eliminated questions applying to other family members and questions relating to only a small proportion of the population.

After reduction, the source questionnaires yielded nearly 900 questions that were possibilities for inclusion in the pretest questionnaire. We wanted to design a questionnaire that could be administered in about 30 minutes. One possible approach to selection of a small number of questions was to sample randomly from among the 900 questions, but this did not appear to be suitable for producing a coherent, workable questionnaire. Our decision was to include questions from each of the four main areas mentioned above as represented in nearly all health surveys.

Within each area we identified a series of core questions. To determine which questions (or which version from one of the groups of high similar questions) to select, we were guided by some general principles. Since we sought to represent the variety of question types found in health surveys, we needed to include both open and closed questions, as well as questions of fact, of perceptions, and of attitudes. We also needed to treat each topic, and to move from topic to topic, in a reasonable manner.

The selection was made from the core questions and others related to the topics selected. It should be emphasized that the apparent "goodness" or "badness" of questions was not a factor in the selection process. As well as we could achieve it, and within the limitations imposed by the overall requirements, the pretest questionnaire included the range of question quality found in health surveys. The final questionnaire consisted of 60 health-related questions as well as four demographic questions needed for analysis.

PLAN OF THE REPORT

Chapter 1 has discussed the need for improved techniques for pretesting survey questionnaires and has described in general terms our research to develop and test new pretesting techniques. Subsequent chapters report this research in detail. Chapter 2 describes and discusses some of the strengths and weaknesses of a typical debriefing session. This provides some background for the remaining chapters. Chapter 3 describes the development and application of behavior coding and special probes. Chapter 4 reports on the use of specially trained interviewers and question ratings. Chapter 5 presents tables of the survey content comparing distributions of responses from the original questions with those from the revised questions.

The final chapter (Chapter 6) consists of a comparative examination and evaluation of the various techniques. The attempt is to generate a cohesive pretest strategy, selecting those techniques that were most effective in identifying and diagnosing question problems. The chapter concludes with a discussion of the implications of the research for pretesting.

An Appendix includes the questions and special probes used in the first and second pretests. It also includes a description of the categories used for the behavior coding, instructions to the coders, and a sample coding form. Also included are a description of the procedures for coding the debriefing, and a sample of the question rating form used by the specially trained interviewers.

REFERENCES

Babbie, E.R.
The Practice of Social Research. Belmont, CA: Wadsworth, 1975.

Backstrom, Charles H., and G.D. Hursch
Survey Research. Evanton, IL: Northwestern University Press, 1963.

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

Cannell, Charles F., and Floyd J. Fowler
A Study of the Reporting of Visits to Doctors in the National Health Survey. Survey Research Center, The University of Michigan. Mimeo, 1963.

____
Comparison of Hospitalization Reporting in Three Survey Procedures. Vital and Health Statistics, Series 2, No. 8, 1965.

Converse, Jean M., and Stanley Presser
Survey Questions: Handcrafting the Standardized Questionnaire. Quantitative Applications in the Social Sciences, No. 63. Beverly Hills, CA: Sage Publications, 1986.

DeMaio, Theresa, J. (ed.)
Approaches to Developing Questionnaires. Statistical Policy Working Paper 10, Office of Information and Regulatory Affairs, Bureau of the Census, 1983.

Groves, Robert M., et al. (eds.)
Telephone Survey Methodology. New York: Wiley, 1988.

Hoinville, Gerald; Robert Jowell, and associcates
Survey Research Practice. London: Heinemann Educational Books, 1978.

Hunt, Shelby D., Richard D. Sparkman, Jr., and James B. Wilcox
The Pretest in Survey Research: Issues and Preliminary Findings. Journal of Marketing Research, 1982:269-273.

Moser, A.C., and G. Kalton
Survey Methods in Social Investigation, 2nd Ed. London: Heinemann, 1971.

Royston, Patricia, and Deborah Bercini
Questionnaire Design Research in a Laboratory Setting: Results of Testing Cancer Risk Factor Questions. Paper presented at the American Statistical Association Convention, Survey Methods Section, 1987.

Schuman, Howard
The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, 31, 1966:218-222.

Selltiz, Claire; Lawrence S. Wrightsman, and Stuart W. Cook
Research Methods in Social Relations, 3rd ed. New York: Holt, Rinehart and Winston, 1976.

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

Figure 1. Study Design

	Technique(s) used		
	Behavior coding	Special probes	Specially trained interviewers
<u>First pretest</u> (with original questionnaire)			
Group 1* (60 respondents, 6 interviewers)	X		
Group 2* (104 respondents, 9 interviewers)		X	
Group 3** (50 respondents, 8 interviewers)	X		X
<u>Second pretest</u> (with revised questionnaire)			
Group 4* (100 respondents, 8 interviewers)	X	X	
Group 5** (50 respondents, 5 interviewers)			X

*Interviews done at the telephone interview facility of the Survey Research Center, The University of Michigan.

**Interviews done at the telephone facility of the Center for Survey Research, University of Massachusetts.

CHAPTER 2

An Evaluation of Interviewer Debriefing in Survey Pretests

Katherine Bischoping

INTRODUCTION

The goal of questionnaire pretesting is to identify aspects of questions that make them difficult for interviewers to ask or for respondents to answer. Since any survey question is likely to present difficulties to some respondent, pretesting also needs to provide estimates of the frequency with which problems occur. Such estimates are required to guide decisions about the need for question revisions.

Many survey organizations use a pretest method that is modelled along the following lines. Five or six interviewers are selected based on their ability to identify problems and willingness to voice their opinions. They are told that the aims of the pretest are to detect problems in the flow and mechanical aspects of the questionnaire, and to produce easily understood questions which convey the intent of study staff. Each interviewer conducts five to ten pretest interviews. During the course of the pretest, interviewers may be allowed to develop and try out question revisions.

Soon after the pretest interviews have been completed, the interviewers, researchers and a discussion moderator (usually an interviewing supervisor) meet for a debriefing. The moderator asks the interviewers for their general impressions of the interviews and then each question in turn is discussed in more detail. On the basis of the comments made at the debriefing, the study staff identify parts of the questionnaire that need revision.

Specific details of pretest procedures vary among survey organizations. While survey methodologists concur that questionnaire pretesting is indispensable, they do not provide consistent recommendations for the procedure. For example, some methodologists advise that especially talented or experienced interviewers be selected for pretesting because they will be most able to identify respondent problems (Converse and Presser, 1986; Fowler, 1984). Others favor including inexperienced interviewers among the pretesters because they will encounter problems that more skilled interviewers are able to avoid (DeMaio, 1983; Hunt et al., 1982). Rarely is it suggested that interviewers be trained thoroughly to develop or enhance their pretesting skills (Belson, 1968).

The recommended ways to select pretest respondents also vary. Hunt et al. (1982) describe two contrasting views: one is that the pretest respondents should be "typical" of the target population; the other is that respondents who are most likely to have problems responding to the questionnaire should be selected for pretesting. The former approach should identify problems that are most likely to occur in production interviews. The latter approach might help identify a larger group of problems but would not give estimates of their prevalence in production interviews.

Yet another issue is whether pretest interviews should be conducted as though they were production interviews or whether they should follow some other procedure. Belson (1981) has found that a large number of respondent problems are not revealed in ordinary interviews, partly because respondents cannot recognize that their interpretation of a question differs from the one intended. Inclusion of probes such as, "Could you tell me what you had in mind when you said X?" or "Would you repeat that question to me, in your own words?" have been used to identify such problems (Cantril and Fried, 1944; Nuckols, 1953; Belson, 1981). In practice these probes are seldom used, perhaps because they tend to disrupt the flow of the interview.

While pretest procedures may use a variety of interviewer and respondent selection methods and various interviewing techniques, the group debriefing of interviewers is common to many procedures. DeMaio (1983) prefers the group debriefing to a series of meetings with individual interviewers because it gives study staff an opportunity to poll all interviewers simultaneously for their views on a possible problem. However the group dynamics literature suggests that the group setting may produce troublesome biases in the expressed opinions of discussion participants. Group members with different experiences may appear to agree because of the strong tendency for groups to reach a consensus (Moscovici, 1985). Members who appear confident and capable may unduly influence consensuses (Kelley and Thibaut, 1969). The behavior of listeners can also affect discussion contributions; an attentive and encouraging audience is associated with contributions of better quality and greater quantity (Rosenthal, 1966). Any of these factors may come into play at a debriefing. For example, interviewers may tend to agree with the most experienced discussants or to refrain from commenting freely in the presence of a defensive study staff. The effects of group dynamics may therefore prevent the study staff from identifying problems and accurately estimating their prevalence.

Cognitive factors may also reduce the usefulness of debriefings. The time lapse between pretest interviews and the debriefing may prevent interviewers from remembering events clearly at the discussion. Comments may also be biased in favor of memories of the most pleasant and unpleasant interactions with respondents (DeMaio, 1983; Tversky and Kahneman, 1982). Furthermore, problems that interviewers can solve easily using routine probes may have low salience to them, and hence not be discussed even though the study staff is interested in identifying all problems.

The preceding issues raise questions about the effectiveness of the group debriefing as a pretest procedure. The focus of this investigation is to evaluate the debriefing as a means of identifying questionnaire problems and estimating their prevalence. Specific issues are: How completely do pretest interviewers inform study staff of the problems they have identified? Can decisions about problem prevalence easily be made on the basis of the debriefing? Does the debriefing give a valid account of interviewer and respondent experiences? How reliable is the debriefing procedure? To explore these issues, we studied two debriefings that were conducted as part of a larger investigation of pretest methodology.

METHODS

Two pretests of a 60-question health questionnaire were conducted by two separate groups of interviewers from the Survey Research Center at The University of Michigan. Respondents were sampled from telephone directories for southeast Michigan.

Group 1 consisted of six interviewers, each of whom conducted ten telephone interviews that were tape recorded with the permission of respondents. The interviewers prepared "comment copies" of the questionnaire, containing suggested revisions, for the study staff to review. They met with the study staff at a debriefing moderated by their supervisor. The discussion was tape recorded for analysis.

The nine Group 2 interviewers used versions of the questionnaire to which a number of special probes had been added to identify problems with questions. The analyses of the Group 2 debriefing exclude from consideration any remarks that interviewers made in reference to these probes. This group conducted a total of 104 interviews. Eight of the Group 2 interviewers prepared comment copies of the questionnaire and seven of them met with study staff and their supervisor for a debriefing that was tape recorded.

Half of the interviewers in each group had not participated in a pretest before and had less than six months interviewing experience. Interviewers were not allowed to deviate from the set question wordings in these pretests, in order that standardized analyses of interviewer and respondent behavior could be conducted. To keep the views expressed at the debriefing as independent as possible, interviewers were asked not to discuss the questionnaire with anyone apart from their supervisor until the debriefing.

The tape recordings of the debriefing discussions were coded following the procedures described in Appendix B.

RESULTS AND DISCUSSION

General Character of the Debriefings

A wide variety of flaws in the questionnaire were discussed at the debriefings. Interviewer problems included difficulty in reading questions exactly as worded, and difficulty in following correct interviewing procedures because of problems with the layout or instructions of the questionnaire. Respondent problems included interrupting question readings with answers, having difficulty in understanding or formulating adequate answers to questions, and being asked questions to which they had supplied answers earlier in the interview. Finally, there were questions that produced negative affect in interviewers or respondents.

When describing respondents' problems with understanding questions and answering adequately, interviewers sometimes referred to concrete signs of respondent difficulties, such as requests for clarification of a question. At other times they gave subjective assessments of respondent difficulties, such as their feelings that a question was not understood. Several comments given at the debriefings went beyond description of question problems and addressed ways in which these problems could be resolved.

The wealth of information given at the debriefings created an initial impression that the pretests were a useful way to discover problems with a questions. Despite the interviewers' lack of specific training in pretesting, they appeared able to identify a wide variety of problems with the questionnaire. Those interviewers who lacked pretesting experience could not be distinguished on the basis of their comments at the debriefing from those with more experience.

Completeness of Reports of Problems

At both debriefings there were large variations in the extent to which interviewers participated in the discussion. One interviewer was talkative to the point that she, rather than the questionnaire, sometimes became the focus of attention. Others made several contributions without dominating the discussion. Still other interviewers gave their general impressions of the questionnaire, commented on the first few questions and lapsed into silence for most of the remaining discussion.

There are several ways that silence at a debriefing can be interpreted. The interviewer's small sample of respondents may have had no difficulties answering a question, or the problems that she noted may have been reported by another interviewer. In either of these cases, silence does not hinder problem identification. However, interviewers may also be silent because they did not notice problems that occurred, did not remember them at the debriefing, felt intimidated, or simply did not feel like making an effort to contribute to the discussion.[1] These explanations of silence reduce the effectiveness of the debriefing as a means of identifying problems.

To find out whether all of the problems interviewers identified were mentioned at the debriefing, statements made at the Group 1 debriefing were compared to the Group 1 interviewers' comment copies. Some critical problems were described in the comment copies, but not in the discussion. For example, one interviewer's comment copy revealed that, when asked whether a risk factor "definitely increases, probably increases, probably does not, or definitely does not increase a person's chances of getting heart disease," respondents answered "definitely" without specifying whether they had an increase or a decrease in mind. This observation was not made at the debriefing. The effectiveness of the debriefing appears to have been reduced by its failure to bring question problems like this to light.

Use of comment copies to evaluate reporting at the debriefing, however, is problematic. The comment copies contained many suggestions that were not

[1] See Diehl and Stroebe (1987) who test these hypotheses in experiments with discussion groups engaged in brainstorming.

accompanied by descriptions of the problems that led to them. Furthermore, the descriptions and suggestions in the comment copies may not be a complete summary of the problems that interviewers encountered in their interviews. In addition, the use of comment copies may have led interviewers to expect the discussion and comment copies to be used in tandem by study staff, and therefore they may not have felt the need to report everything at the debriefing. Thus, while this investigation uncovered a potential weakness in the debriefing procedure, the study design did not allow this issue to be examined rigorously.

Ease of Estimating Problem Prevalence

There are several difficulties in using the debriefing discussion to judge the prevalence of problems. First of all, it was uncommon for all interviewers to speak about any given question; in consequence, the number of respondents whose experiences were described varied across questions. Those interviewers who did speak about a particular question frequently addressed different issues. Thus an accurate estimate of the number of respondents who had a specific problem with a question could not be obtained.

Efforts to estimate problem prevalence were further complicated by the language interviewers used to describe problem frequency. Instead of giving the number of respondents who had problems with a question, interviewers used vague quantifiers such as "some" and "many", which vary in meaning among individuals and across contexts (DeMaio, 1983; Bradburn and Miles, 1979). Many of the comments made at the debriefing were even more vague about problem frequency. Sometimes it was not clear whether interviewers were describing problems that actually had occurred in pretest interviews or whether they were speculating about problems that could conceivably occur. For example, an interviewer asking, "What's a health professional?" might mean that a number of respondents had not understood the term or that she herself found the term confusing. The issue was clouded further by interviewers' frequent agreements with one another. Agreement could mean that the interviewers had similar experiences or merely that their hypotheses about the question concurred. Finally, it was difficult to interpret the vivid examples of individual respondents' odd reactions to questions since it was not clear whether these were meant to entertain the group or to point out important problems with the questionnaire. Although all of these types of comments contain some information about problem prevalence, it is impossible to determine their significance.

In sum, the analysis of the debriefing discussions showed that they could not be used to produce precise estimates of the number of respondents who had a particular problem with a question. The less ambitious objective of judging whether a substantial number of respondents had problems with a question was complicated by the qualitative and frequently ambiguous nature of the information that interviewers gave.

Accuracy of Reports of Interviewing Experiences

To study how accurately the interviewers' described their pretesting experiences, the Group 1 discussion was compared with the results of an analysis of interviewer and respondent behavior that had been done for the Group 1 interviews (Chapter 3). For each of the 60 questions in the questionnaire, a coding system was applied that measured the percentage of interviews in which certain problem-indicating behaviors occurred. The behaviors included:

READ - the interviewer made a slight or major error in reading the question.

INT - the respondent interrupted the initial reading of the question.

RESP - the respondent requested repetition or clarification of the question, qualified a response to indicate uncertainty, gave an inadequate or "don't know" answer, or refused to answer.

Another coding scheme was developed in order to assess the evidence interviewers gave at the debriefings about the same set of behaviors (see Appendix B for details). For each question the presence or absence of the following information in each of the debriefing discussions was coded:

READ - there is some mention of a reading problem, or a change is suggested to improve ease of reading.

INT - there is some mention that respondents interrupt with answers or a suggested change to alleviate this problem.

RESP - interviewers' comments indicate that a substantial number of respondents requested repetition or clarification of the question, qualified their responses to indicate uncertainty, gave inadequate or "don't know" answers, or refused to answer the question.

By comparing the results of the behavior analysis and the Group 1 discussion coding it was possible to see how well the Group 1 interviewers' descriptions of their experiences corresponded to their actual experiences for a limited set of behaviors that are indicators of question problems. Other issues that interviewers discussed, such as layout problems, were not measured by the behavior coding system.

The Group 1 discussion identified ten questions that interviewers found difficult to read exactly as worded. This set of questions was compared to the set of ten questions that had the highest levels of reading errors according to the behavior analysis. Ideally these two sets of questions would be identical -- yet they matched in only three of the ten cases. A similar comparison was made for the set of five questions for which interruptions were noted at the debriefing. For four of these five questions the debriefing and behavior analysis results were in agreement. Of the 34 questions for which interviewers described the remaining set of respondent problems, 26 were among the 34 questions with the highest levels of problems according to the behavior analysis.

These comparisons indicate that the Group 1 interviewers' statements about reading problems led to very inaccurate identification of questions with high levels of reading errors. The debriefing led to more accurate identification of questions that were frequently interrupted or showed respondent problems. However, the presence of discrepancies between the discussion and the behavior coding, particularly for respondent problems, is cause for concern.

Sources of differences between behavior analysis and discussion coding results

To investigate possible causes for the differences between the discussion coding and behavior analysis, we consulted the tape-recording of the debriefing discussion, the notes taken during the discussion coding, and the results of the behavior analysis, for those questions that had inconsistent results. Some of the disagreements between the behavior analysis results and the discussion coding may of course be attributed to random error, incorrect application of either coding scheme, or lack of correspondence between the behaviors measured by the two methods.

One group of questions that was investigated had high levels of reading errors, interruptions, or other respondent problems according to the behavior analysis but the debriefing coding did not indicate significant levels of these problems. Several explanations for this type of disagreement seemed plausible. For some questions, a high level of problem-indicating behavior codes may have been produced by a combination of several small problems that interviewers were unable to detect. For other questions, interviewers mentioned problems at the debriefing in a way that was not clear or compelling enough for the discussion coding to register that a problem was present. A third explanation is that interviewers may not always have realized that a respondent behavior measured for the behavior analysis could be interpreted as a sign of a problem. For example, interviewers may have ignored instances when respondents qualified their responses to indicate uncertainty because this behavior did not disrupt the interview.

Possible explanations related to the group dynamics at the debriefing also arose from the analyses. It appeared that discussion of other problems with a question may have caused interviewers to forget problems they wanted to mention, or to refrain from mentioning them because they would be solved by proposed revisions of the question. As well, time pressures in the debriefing may have prevented systematic discussion of problems with some questions. In a few cases, interviewers seemed eager to describe problems with a question following the one under discussion and moved ahead without the authorization of the moderator.

Another group of questions that was studied were identified as problematic by the debriefing but did not have high levels of problem-indicating behavior codes. A different set of explanations for these differences is proposed. First, interviewers may have mentioned problems that were salient to them but would not result in high behavior code levels. An interviewer's comment that she stumbled over an awkwardly worded question the first time she read it would fall into this category. Second, the reliability of the discussion coding scheme appeared poor in a few cases -- questions were sometimes judged to be problematic on the basis of weak evidence. For other questions, comments about a series of related questions were given together so that the better questions in the series could not be distinguished from the worse. Finally, several of the questions for which the debriefing codes and behavior analysis results disagreed were not asked of all respondents. Since this disagreement persisted when behavior code levels for each question were calculated using only interviews in which it was asked, the source of the problem may be that interviewers generalize less well from small sample sizes than large ones.

These observations suggest that much of the disagreement between the discussion and the behavior analysis results may arise from factors that could be controlled, such as the way that the debriefing discussion is moderated, the training that interviewers receive, and the sample sizes that are used. These points will be addressed in the discussion.

Reliability of the debriefings

In order to assess the reliability of debriefing sessions for identifying questions that are sources of interviewer or respondent problems, the Group 1 and Group 2 debriefings were compared. In addition to the codings of READ, INT, and RESP problems described above, the debriefings were coded for the presence or absence of the following problems with each question:

INST - the interviewer instructions are not clear, or the layout of the questionnaire is difficult to follow.

INAP - the respondent has supplied the required information earlier in the interview, or given information that makes it clear that the question is inappropriate.

MOOD - the interviewer or respondent finds the question embarrassing, silly, or otherwise upsetting.

The agreement between the Group 1 and Group 2 discussions is presented in Table 1.

The kappa statistic was used to evaluate the degree of agreement beyond chance in each of the six problem categories. The agreement between the two groups regarding READ problems was notably poor (K=0.05). Although individual differences in question reading skills and ability to note reading errors are to be expected, it is hoped that the sample of interviewers participating in a pretest is large enough to allow common reading problems to be detected. Evidently this was not the case. In contrast the agreement for INT problems was perfect (k=1.00) and the agreement for INAP problems was very high (k=.81), suggesting these may be more memorable difficulties. For the remaining categories of problems, including RESP problems, the agreement beyond chance was moderately good (k ranging from 0.41 to 0.59).

The Group 2 debriefing gave evidence of RESP problems for more questions than the Group 1 debriefing did. This difference may be explained to some degree by the slightly larger samples of interviewers and respondents for Group 2. However, Group 2 had a larger number of extremely quiet interviewers than Group 1, which would tend to balance the effect of sample size difference. It is also possible that the Group 2 interviewers were sensitized to respondent problems because of the special probes that were included in their questionnaires, or that the attempt to exclude interviewer observations based on the special probes from analysis failed.[2]

Although the agreement between Group 1 and Group 2 interviewers about RESP problems is moderately good, this assessment may be misleading. When the nature of the respondent problems that the two groups identified is compared, further differences emerge. For example, for the question, "As close as you can remember, what was your blood pressure, in numbers?" evidence was given at both debriefings that some respondents didn't know the answer. Group 1 interviewers noted that some respondents gave the two numbers making up the blood pressure reading in the incorrect order, while Group 2 interviewers said respondents sometimes answered, "it was normal," or "it was high". Another question that illustrates the differences in the groups' reports is: "About how many days a week do you think a person needs to exercise, to strengthen the heart and lungs?". The Group 1 interviewers stated that respondents felt uninformed about this issue and sometimes gave the inadequate answer "every other day". The Group 2 discussion indicated that respondents didn't really understand the question; respondents asked what "strengthen" meant and did not appreciate that the question referred to an average person. In cases like these, the two debriefings would lead to different revisions of the question or the response categories, and neither might resolve the full range of problems discussed by the two groups of interviewers.

CONCLUSION

Interviewer comments at the two debriefings gave the initial impression that the procedure is an effective means of identifying a variety of problems with questions. However, a closer examination of the debriefing revealed a number of potential problems. Interviewers did not inform the study staff of all the problems they found with the questionnaire at the debriefing (although their comment copies could be used to obtain additional information). Moreover it was difficult to estimate the prevalence of those problems that were discussed -- interviewers did not all give information about the frequency of each problem, used vague terms to describe frequencies, and did not distinguish between their conjectures and facts.

Judgments that were made about problem prevalence on the basis of the debriefing did not always correspond to the interviewing experiences, as measured by behavior analyses of the interviews. In particular, for question reading problems, poor agreement was found. We hypothesize that, in part, inconsistencies arose from unorganized or incomplete discussion of questions, failure of the interviewers to recognize symptoms of problems that study staff thought important, and the difficulty of accurately reporting the experiences of small samples of respondents. In addition, while the reliability of the

[2] As part of the pretest methodology study, a third pretest was conducted in which special probes were used. At the debriefing for this pretest interviewers did not appear to be especially sensitive to respondent problems. This suggests that individual differences, rather than a reaction to the use of special probes, may have caused the Group 2 interviewers to be more sensitive to respondent problems than the Group 1 interviewers.

debriefings in identifying questions with various types of problems was fairly good, the reliability for question reading problems was poor. Even when both groups agreed that a respondent problem was present, they did not always agree on the nature of the problem. The existence of differences between the two pretests implies that the group debriefing procedure may not identify all problems of a questionnaire, may not produce good estimates of their prevalence, and may lead to question revisions that fail to resolve the most frequently occurring interviewer and respondent difficulties.

It is difficult to determine the extent to which these conclusions apply to other group debriefings. One reason is that the data used come from only two pretests. A second reason is that certain findings are dependent on the discussion coding system that was used in this study. The coding entailed time-consuming estimation of the amount of information interviewers gave about a specific group of respondent behaviors -- other researchers making judgments on the basis of a debriefing might work more rapidly, take a wider assortment of evidence into account, evaluate the relevance and potential accuracy of individual contributions, or rely more heavily on their intuition. Despite these limitations, this research points to weaknesses in the group debriefing procedure that may often be present.

The findings from this study suggest approaches that could be taken to improve the problem identification and prevalence estimation in a group debriefing. In particular, recommendations are made regarding pretest interviewer training, debriefing moderator training, and the sources of information used in pretesting.

Improvements in pretest interviewer training might be used to alleviate a variety of potential weaknesses of the group debriefing. For example, interviewers may not always recognize indicators of problems with questions. This weakness might be addressed by giving interviewers descriptions of respondent behaviors that can be interpreted as signs of problems (including, for example, expressions of uncertainty about the accuracy of a response). A more thorough training program could easily be devised. For example, supervisors could prepare a tape-recording of an interview that contains a variety of problems. By comparing the list of problems occurring in this interview to the lists of problems that pretest interviewers identify, supervisors would be able to determine which problems are overlooked by each interviewer and give further training accordingly.

Two other weaknesses of the debriefing could be addressed by interviewer training. These are the possibility that interviewers make incomplete reports of problems they identify, and difficulties in making judgments about problem prevalence using the debriefing discussion. These problems might be reduced by ensuring that pretest interviewer training conveys the study staff's objectives clearly. .If the study staff wants to identify problems successfully, then training should emphasize that all of the problems that are noted should be reported at the debriefing. If the study staff wants to estimate problem prevalence, then interviewers should be asked to make a practice of distinguishing between their conjectures about possible respondent problems and the actual experiences of respondents. If the study staff wants to determine whether question objectives are communicated to respondents, then interviewers should be told what the question objectives are. By these means, the benefits of the opportunity for information exchange provided by the debriefing may be realized more fully.

The debriefing moderator should also be given instructions that will lead to improvements in the debriefing procedure. First of all, the moderator should be instructed to focus the content of the discussion so that it meets the study staff's requirements and reinforces the pretest interviewers' training. For example, s/he could request that interviewers specify whether their comments pertain to respondents' experiences or are speculations about problems that they might have if they were respondents. The moderator could also be instructed in ways to deal with disruptions in the question-by-question discussion. For example, we hypothesize that simultaneous discussion of a series of related questions causes question problems to be exaggerated. To prevent this, the moderator could ask interviewers for their general impressions of questions in the series and then request discussion of each question in turn. This might encourage a more complete description of problems and permit comparison of the relative merits of questions in a series.

Besides having some control over the content and pace of the debriefing discussion, the moderator can act to control participation at the debriefing by directly calling on individual interviewers or by using indirect tactics to encourage silent interviewers and to discourage overly talkative ones.[3] In light of the difficulties of interpreting interviewer silence at the debriefing, it might appear useful to recommend that the moderator exercise this authority and ensure that all interviewers participate in the discussion of each question. However, were this done, interviewers might feel considerable pressures to agree with the opinions of apparently competent group members. The comments made at the debriefing might then be as ambiguous as the silence we observed.[4] For this reason, we are reluctant to recommend that the discussion moderator attempt to elicit comments from all interviewers about each question problem.

Regardless of the efforts of the discussion moderator and the interviewers participating in the group debriefing, cognitive factors alone may limit the value of this pretest procedure. As Belson (1981) has found, respondents are often not aware that they are misinterpreting questions. In addition, DeMaio (1983) points out that interviewers may be unaware that they are reading questions incorrectly. Interviewers also may not be able to remember the frequency of events clearly enough to avoid using vague quantifiers. Finally, the small numbers of interviewers and respondents

[3] Merton et al. (1952) describe ways in which a discussion moderator can reduce the input of talkative group members and encourage comments from quiet members.

[4] Yoell (1974) contends that a discussion group leader cannot control the group dynamics completely. He also gives compelling evidence that comments made in a group discussion can be motivated by speakers' feelings about other participants, or by a desire to say what the moderator wants to hear.

involved in pretest interviews provide a poor basis from which to make judgments about the larger populations that will participate in production interviews.

Unavoidable weaknesses in the group debriefing procedure could be dealt with most successfully by using alternative or additional methods to strengthen the pretest. For example, instead of using a debriefing to identify reading problems, these problems be identified by noting the reading errors of a large group of interviewers asked only to read through a questionnaire. In order to obtain complete information about other problems, methods that produce complementary information should be combined. For example, the diagnostic probes that Belson (1981) describes could be used to identify problems of question misinterpretation that are not brought to light in ordinary interviews. Three other techniques that could complement the debriefing by producing precise prevalence estimates are interviewer questionnaires about specific question problems that are completed after each pretest interview (DeMaio, 1983), interviewer rating scales that are completed after all pretest interviews (Chapter 4), and the behavior analysis system used in this study (Chapter 3). If prevalence estimates from one of these methods are available by the time of the debriefing, the discussion could be focused on identifying the nature of problems with questions that have a high frequency of problems. These examples show how an integrated set of pretest procedures could be used to provide richer information than the group debriefing produces alone.

To some extent, pretesting must always be a matter of "hypothesis testing by hunch and judgment" (Converse and Presser,1986). It is our hope that the recommendations given in this study for interviewer training, moderator behavior, and use of additional pretest procedures will lead to improvements in the hunches and judgments of researchers using pretests.

REFERENCES

Belson, William A.
Respondent understanding of survey questions. Polls 3 (4), 1968.

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

Bradburn, Norman M. and Carrie Miles
Vague quantifiers. Public Opinion Quarterly, 43, 1979:92-101.

Cantril, Hadley and Edrita Fried
The meaning of questions. In Hadley Cantril (ed.), Gauging Public Opinion. Princeton: Princeton University Press, 1944.

Converse, Jean M., and Stanley Presser
Survey Questions: Handcrafting the Standardized Questionnaire. Quantitative Applications in the Social Sciences No. 63. Beverly Hills, CA: Sage Publications, 1986.

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

Diehl, Michael and Wolfgang Stroebe
Productivity loss in brainstorming groups: Toward the solution of a riddle. Journal of Personality and Social Psychology 53 (3), 1987:497-509.

Fowler, Floyd J.
Survey Research Methods. Beverly Hills, CA: Sage Publications, 1984.

Hunt, Shelby D., Richard D. Sparkman, and James B. Wilcox
The pretest in survey research: Issues and preliminary findings. Journal of Marketing Research 19, 1982:269-273.

Kelley, Harold H. and John W. Thibaut
Group problem solving. In Gardner Lindzey and Elliot Aronson (eds.), The Handbook of Social Psychology, Vol.IV. Menlo Park, CA: Addison-Wesley, 1969.

Merton, Robert K., Marjorie Fiske, and Patricia Kendall
The Focussed Interview. New York: Bureau of Applied Social Research, Columbia University, 1952.

Moscovici, Serge
Social influence and conformity. In Gardner Lindzey and Elliot Aronson (eds.), The Handbook of Social Psychology, Vol.II. New York: Random House, 1985.

Nuckols, Robert C.
A note on pre-testing public opinion questions. The Journal of Applied Psychology 37 (2), 1953:119-120.

Rosenthal, Robert
Experimenter Effects in Behavioral Research. New York: Meredith, 1966.

Tversky, Amos and Daniel Kahneman
Judgment under uncertainty: Heuristics and biases. In Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982.

Yoell, William A.
How useful is focus group interviewing? Not very...post-interviews reveal. Marketing Review 29, 1974:15-19.

Table 1. Comparison of Group 1 and Group 2 discussion codes for interviewer and respondent problems.

READ

Number	Group 2:	
of Qs	absent	present
Group 1:		
absent	46	3
present	9	1

k=.05

INT

Number	Group 2:	
of Qs	absent	present
Group 1:		
absent	55	0
present	0	4

k=1.00

RESP

Number	Group 2:	
of Qs	absent	present
Group 1:		
absent	16	10
present	3	30

k=.54

INST

Number	Group 2:	
of Qs	absent	present
Group 1:		
absent	40	5
present	7	7

k=.41

INAP

Number	Group 2:	
of Qs	absent	present
Group 1:		
absent	52	1
present	1	5

k=.81

MOOD

Number	Group 2	
of Qs	absent	present
Group 1:		
absent	45	4
present	3	7

k=.59

Note: Fifty-nine questions were used in this comparison. The sixtieth was excluded because of the possibility that the tape recording of the relevant discussion was incomplete for Group 2. For this question the two groups appeared to have inconsistent INT codings.

CHAPTER 3

New Methods for Pretesting Survey Questionnaires

Lois Oksenberg, Charles Cannell, and
Graham Kalton, Survey Research Center,
The University of Michigan

INTRODUCTION

Numerous methodological investigations have demonstrated that poorly designed questions make major contributions to survey error. Some problems with survey questions arise from difficult wording or complex phrasing, some problems arise when questions fail to communicate unambiguously to the respondent what is wanted, and others arise when the information requested places too great a burden on respondents to recall information or to organize their responses.

It is crucial to the quality of a survey to identify problems and improve questions prior to data collection. Commonly used pretest practices frequently fail to accomplish the task and, as a result, many problem questions filter through pretests into final survey questionnaires.

This paper describes the development and testing of two methods designed to provide objective information on the existence of question problems and their causes. The paper also illustrates how the methods can be used during pretesting to identify problems and diagnose their sources, and the use of such information to improve questions.

The methods are systematic analysis of respondent and interviewer behaviors in the question-and-answer process, and use of special follow-up probe questions to investigate respondents' understandings of questions and response difficulties. Neither of these techniques is new, both having been used previously in methodological investigations. The purpose of this research is to adapt and develop the techniques for efficient use in improving survey questionnaires.

The analysis of behavior is frequently used in social psychology as a vehicle to illuminate social and psychological processes. Observation of behavior can be used not only to categorize types of activities but often also to learn about the dynamics and determinants of behaviors. The main purpose of the research reported here was not to study theoretical issues about interaction, but rather to learn about the nature and prevalence of communication problems that exist when interviewers and respondents are confronted with survey questions. The interview interactions are used as data that might indicate when and why interviewers or respondents were having difficulty with questions.

Previous research (e.g., N.C.H.S., Vital and Health Statistics, Series 2, No. 109, forthcoming; Morton-Williams and Sykes, 1984) has shown that several types of behaviors indicate that interviewers or respondents are having trouble with a survey question. Wording changes by interviewers often indicate a question that is difficult to ask as printed. Respondent requests for clarification or repeats of the question can demonstrate comprehension difficulties. Inadequate answers (answers that do not meet question objectives) indicate some sort of response difficulty, such as problems with recalling and organizing information needed to answer or inadequate understanding of the question.

Problem-indicating behaviors can be captured and quantified through a system for coding interview behavior. Such systems have been used by a number of investigators, at first concentrating on interviewer behavior for monitoring and evaluating interviewer performance (Cannell, Lawson and Hausser, 1977). Methods subsequently were developed to include respondent behavior to investigate the question-answer process more generally (Cannell and Robison, 1971; Dijkstra, 1985; Morton-Williams, 1979; Marquis, 1969; Marquis, 1971a, 1971b). The current research adapted and developed those codes of interviewer and respondent behaviors to identify some type of problem in asking or responding to survey questions.

A second method to identify respondent difficulties is the use of special follow-up probes. Information from the probes may supplement behavior coding or be used to reveal problems not evident in the response behavior. Investigations by Cantril and Fried (1944), Schuman (1966), Belson (1981) and others have demonstrated the value of this technique to explore question interpretation.

Some problems respondents have with questions are obvious in their behavior, while others are hidden. Any of the following situations can occur when a question is asked:

1. The respondent understands the question correctly and provides an answer.

2. The respondent misinterprets the question and provides an answer.

3. The respondent does not understand the question, seeks clarification, and provides an answer.

4. The respondent may not understand the question, but may feel that to acknowledge difficulties communicates a personal inadequacy, or may be unwilling to make the effort to ask for clarification. He or she therefore provides an answer.

If the answer is adequate (that is, can be coded according to question objectives) the comprehension problems in the second and fourth scenarios will not be revealed in behavior. This may be particularly likely to occur with closed questions, for which the respondent is asked to choose among a set of alternative answers. Likewise, for any of the scenarios, answers that can be coded according to question objectives may mask inadequate recall and organization of information needed for accurate responding. Special probes are potentially useful for revealing such comprehension and answer formation problems. To provide a range of experience of the value of special probes in pretests, a variety of probes were designed and tested on a number of survey questions.

To be most useful, pretest techniques need to do more than identify problematic questions; they need to identify the sources and nature of the difficulties to serve as a basis for question improvement. The behavior coding and special probe results provide data from which one can generate hypotheses as to the most likely sources of problems for a question, and the question can then be revised.

General Design

This research was based on questions from existing health surveys. We collected many health survey questionnaires from governmental agencies, universities, and private organizations. These questionnaires covered the main subject matter areas included in health surveys. Sixty questions were selected, covering such topics as: visits for medical and dental care; health care plan membership; various aspects of health status; and nutrition, exercise, cancer and heart disease risks, and knowledge about AIDS.

Figure 1 gives the main components of the study reported here. The original questionnaire was administered to two groups in the "first pretest." Group 1 interviews were coded using behavior codes. For Group 2 interviews special probes were included during the interview and at the end. Questions identified as problematic in the first "pretest" were revised. The revised questionnaire was administered in the "second pretest," in which special probes and behavior coding were applied to the same interviews (Group 4). We assessed question improvement by comparing behavior coding results for Group 1 and Group 4. Although Group 4 interviews included several special probe questions, we judged that those used during the interview had the appearance of normal survey questions that fitted naturally with the flow of the questionnaire. As a result, these special probe questions were unlikely to have affected respondent behavior with regard to other questions.

This research was part of a larger study of pretesting techniques. While the other components generally are not relevant to the techniques explored here, we did make use of some findings from another group of interviews (Group 3). This will be discussed later.

Respondents were residents of southeastern Michigan selected from computerized lists based on telephone directories. Interviewing was done by telephone and all interviews were tape recorded (with the permission of the respondent). Coding of interview behavior was based on these recordings.

The paper has three parts. The first part reports research based on the coding and analysis of interview behavior. The second part reports research on the use of special probe questions. The third part reviews the strengths and weaknesses of the techniques and discusses how they may be employed in combination.

ANALYSIS OF INTERVIEW BEHAVIOR

Development and Application of the Behavior Coding System

This section describes the use of behavior coding to identify interviewer and respondent difficulties with questions, and demonstrates that this can be done simply and reliably. The behavior coding system was developed with the aim of capturing those components of the interviewer-respondent interaction that indicate problems with questions. The codes were directed at obtaining information that would help identify the sources of the problems. While this research used a detailed coding scheme, the aim was to identify a few, easy to apply codes that would capture problems when they arise and that could be readily applied in regular pretesting. We sought to develop a procedure in which coding would keep pace with the interview and which facilitated aggregation of coding results across interviews to provide indicators of problems.

From the beginning it was apparent that coding all interviewer and respondent behavior was too time-consuming. Coders would not be able to keep up with the interview. In addition, coding both sides of the interaction appeared unnecessary, since interviewer behavior tends to be reactive to the behavior exhibited by the respondent. For example, if the respondent gives an incomplete answer, the interviewer asks for more information. If the respondent asks for the meaning of a term, the interviewer typically reacts to the inquiry, etc. This simplified our technique since it meant that coding only the respondent behavior was sufficient to identify problems.

The only interviewer behavior that was coded was accuracy and completeness with which the question was initially asked. Question reading was categorized into exact readings, readings with slight wording changes that did not alter meaning, readings with major wording changes, and readings that were broken off before the question was completed.

Respondent behavior categories were: interrupting question reading with an answer, seeking clarification or a repeat of the question, giving an adequate answer, giving a qualified answer (an adequate answer for which the respondent indicated uncertainty about its accuracy or completeness), giving an inadequate answer, giving a "don't know" answer, and refusing to answer. In this context adequacy merely meant that the answer could be coded according to question objectives. In the absence of additional information, the accuracy or completeness of answers could not be evaluated.

Figure 2 gives a brief description of the codes. The coding system involved categorizing the relevant interviewer and respondent behaviors for each question asked in each interview. Codes were assigned for both respondents' initial and subsequent response behaviors to a question. 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 each response was coded. For example, consider respondent behavior for a question that begins with an inadequate answer (Code 7) followed by a request for clarification (Code 3), all in the respondent's first speaking turn. Following the intervening interviewer behavior of providing clarification, the respondent gives another inadequate answer (Code 7) followed by an adequate answer (Code 5). This sequence of respondent behavior would be coded "7,3" for the first turn at speaking and "7,5" for the second turn.

Coding Procedures, Coders and Coder Training

Three experienced telephone interviewers were employed as coders for the first pretest interviews. Each coder coded approximately equal numbers of interviews from each interviewer. Two of them also coded the second pretest interviews. Since the coders were well familiar with interviewing techniques, their training as coders was efficiently accomplished in a few hours.

Indices of Question Problems

Many different indicators of problems with questions can be derived from the behavior codes. We examined a number of indicators and their interrelationships before selecting the final indicators. As with the design of the code categories, the goal here was to devise indicators that would adequately and efficiently identify interviewer and respondent problems.

We investigated three main approaches for respondent behavior. The first was to consider only the first behavior coded for the question, rather than all behaviors for the question, on the presumption that the first reaction to a question was the most likely to reveal problems. However, our analysis demonstrated that respondents sometimes gave adequate answers followed later by a problem indication. Since this situation would not be identified by coding only first reactions, an analysis of all behavior for the question was preferred.

The second approach considered was to base problem indicators on the number of times the relevant type of behavior was coded. For example, if a respondent gave three inadequate answers in the course of answering a question, the score on the inadequate answer indicator would be 3. However, analysis of the behavior coding made clear that multiple instances of a code were infrequent, and that the number of times the behavior occurred was very highly correlated with whether or not the type of behavior occurred at all. In addition, a single example of the behavior indicated a problem with the question, and additional behaviors add little useful information.

On this basis we adopted the third approach, which was to base the indicators on whether the relevant type of behavior occurred at all. With this approach a respondent who gave one or more inadequate answers (as in the previous example) would receive a score of 1 on the indicator, while a respondent with no inadequate answers would receive a score of 0. The following indicators were chosen to tap the range of problems that behavior coding might identify. Values of the Kappa statistic, a form of intraclass correlation coefficient used as a measure of intercoder agreement, are given for the indicators. All but one of the kappa values represent good or excellent agreement among coders.[1]

1. Slight changes. Interviewer read the question with slight changes. "slight changes"). (kappa=0.73)

2. Major changes. Interviewer read the question with major changes, or did not complete the question reading.[2] (kappa=0.72)

3. Interruption. Respondent interrupted the question reading with an answer. (kappa=0.90)

4. Clarification. Respondent requested clarification, explanation, or repeat of the question. (kappa=0.93)

5. Qualified answer. Respondent gave a qualified answer. (kappa=0.56)

6. Inadequate answer. Respondent gave an inadequate answer (inadequate answer). (kappa=0.85)

7. "Don't know." Respondent gave a "don't know" answer. (kappa=0.86)

8. Refusal. Respondent refused to answer. (Kappa was not calculated since the code was assigned only a few times in the study.)

9. Respondent problem. Respondent behavior indicated some problem, that is, one or more of the respondent problem indicators (numbered 3 through 8 in this list) was scored 1. (kappa=0.88)

10. No adequate answer. Respondent never gave an adequate answer. (kappa=0.75)

The last two indicators -- "respondent problem" and "no adequate answer" -- were designed as summary indicators of some sort of respondent problems. While not useful in diagnosing sources of trouble, these measures might detect

[1] To assess the level of agreement among coders, approximately one out of every 10 interviews was coded independently by a member of the study staff who had been centrally involved in developing the coding system and in training the coders. In all, 19 interviews (13 first pretest and 6 second pretest interviews) involving 1098 question askings were independently coded. Scores on all 10 indicators were calculated for each question, once based on the coding from the regular coder, and once based on the coding from the staff member. Kappa for all but one of the indicators represent excellent or very good agreement. The exception, whether or not the respondent gives a qualified adequate answer, was only fair to good.

[2] Question reading with major wording changes and breakoffs in reading (codes M and B in Figure 2) were combined and treated as the same type of behavior.

a cumulation of small problems, and might be useful in a preliminary screening to identify which questions should be examined further for respondent problems.

Several possible additions to the ten indices were also considered. These included measures based on whether or not the respondent's first behavior for a question was an adequate response, the number of respondent turns at speaking for the question, and the total number of respondent behaviors for the question. Ideally, respondents should provide an adequate answer with no extra behavior. Failure to follow this pattern may indicate the respondent is having trouble with the question. However, an examination of these measures showed that they provided little information about question problems beyond that available from the indicators we selected. For example, scores for respondents that showed whether or not their first behavior for a question was an adequate response were extremely highly correlated with scores on the "respondent problem" indicator (9). Furthermore, when the few open questions in the questionnaire were excluded, both the number of turns and the number of respondent behaviors were highly correlated with scores on the same indicators.[3] Neither the adequacy of the first response nor the sheer quantity of behavior adds to information available from the summary "respondent problem" indicator.

In order to use behavior coding to learn about questions, there needs to be assurance that coding results reflect aspects of the question and not characteristics merely of the particular interviewers or particular samples of respondents. To investigate the stability of the behavior coding, assignment of problem indicators in Group 1 was compared to that from Group 3. These groups included different respondents, interviewed by different interviewers who were trained and supervised by different organizations. Although Group 3 interviewers had been specially trained to recognize problems with questions, they, as well as Group 1 interviewers, used standard interviewing techniques and procedures. The results showed that levels on the problem indicators for questions were quite comparable for the two groups. That is, the same questions were identified by the same indicators. The one difference of note was that one group had substantially more slight changes in wording in reading the questions than did the other, probably reflecting differences in training. These findings demonstrate that coding results for respondent behaviors are robust under different interviewing conditions and can be used to evaluate questions without concern that results would be substantially different with different sets of interviewers or respondents.

[3] The small number of open questions were excluded from this calculation because interviewers routinely used "anything else" type probes with them, thereby automatically increasing the amount of respondent behavior for these questions.

Results and Discussion

Identification of problematic questions

We analyzed behavior coding results for each question in the Group 1 interviews. For each problem indicator we calculated the percentage of interviews in which the indicator was assigned among those interviews in which the question was asked. Table 1 displays the incidence of problems for the 60 questions in the original questionnaire.

The table shows that a sizeable number of questions had high levels of slight changes in question wording. For example, 18 of the 60 questions had slight wording changes in 15 percent or more of the interviews. In contrast, major wording changes were relatively rare. Such changes often are inadvertent, but sometimes interviewers intentionally change question wording in an attempt to improve the question. Under the coding rules employed, major changes also include times when the interviewer discontinued reading the question because the respondent interrupted.

As Table 1 shows, respondents only infrequently interrupted question reading to give an answer. However, for a quarter of the questions, 15 percent or more of the respondents requested clarification or repeat of the question. Inadequate answers were the most frequently occurring problem indicator for the questionnaire. For over two-thirds of the questions, 15 percent or more of the respondents gave inadequate answers. Qualified answers were less common. "Don't know" answers were rather uncommon and refusals to answer were practically nonexistent.

The information provided by the behavior codes can best be illustrated by looking at some examples. The questions chosen illustrate different kinds of problems that are identified by behavior coding. Behavior coding results from Group 1 for these examples are given in Table 2.

Ex. 1. "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?"

Ex. 2. "How much did you pay, or will you have to pay, out of pocket for your most recent visit? Do not include what insurance has paid for or will pay for. If you don't know the exact amount, please give me your best estimate."

As Table 2 shows, these questions have high rates of respondent interruptions. Respondents answered these questions before the interviewer had completed asking them in many of the interviews. These questions probably appeared to respondents to be completed before the question reading had been in fact completed.

Ex. 3. "What do you think are the warning signs or symptoms of cancer?" Table 2 shows that about a quarter of the respondents requested clarification, about a fifth gave inadequate answers, and four-tenths gave "don't know" answers to this question. These figures make it clear that the question is causing considerable difficulty. There appear to be several possible kinds of problems, including comprehension difficulties or difficulties with the response task.

The next example was the third in a set of questions prefaced by, "We're interested in how well people take care of themselves. Do you think you do very well, fairly well, or not so well as far as...?" The third question read:

Ex. 4. "taking care of your teeth or dentures?"

Even though interviewers were allowed to repeat the response choices, Table 2 shows that over a quarter of the respondents gave inadequate answers. It seems likely that these respondents did not know or remember what form the response was to take, or did not appreciate the importance of responding in terms of the categories.

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

Interviewers were required to record the month and year. Nearly nine out of ten respondents gave inadequate answers, and over a quarter gave qualified answers. From an examination of the question, it seems likely that some of the difficulty occurred because the response form was not clearly specified. It is also likely that in some cases recalling the exact month and year is a difficult task for the respondent.

Ex. 6. "Would you say that you are physically more active, less active, or about as active as other persons your age?"

Over a fifth of the respondents qualified their responses to show uncertainty in answering this question. It appears likely that the task of comparing their health to that of others their age was difficult.

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

This question was intended to refer to dental care, which was the topic of the preceding question. Nearly a third of the respondents requested clarification, and about the same proportion gave inadequate answers to the question. One likely reason for the requests for clarification was that respondents did not know what kinds of treatments or examinations were intended. And, like Example 5, the question did not specify the response form. Since the categories the interviewer was to use to record the answer were not stated, the high levels of inadequate answers are not surprising.

Ex. 8. "I am going to read a list of things which may or may not affect a person's chances of getting <u>heart disease</u>. 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 cholesterol?"

Even though interviewers could repeat the response categories when asking the questions, all the questions in this set had high levels of inadequate answers (all above 50 percent). As Table 2 shows, 77 percent of respondents gave inadequate answers to question 8b. Respondents did not answer in terms of the categories provided. As with Example 4, they may have had difficulty remembering the response categories, or did not understand the importance of responding in those terms.

The next example followed a question about consumption of butter.

Ex. 9. "What is the number of servings on a typical day?"

Seventeen percent of the respondents requested clarification for this question. They may have been confused about what food(s) the servings referred to.

Diagnoses of Problem Sources and Question Revision

The examples illustrate the information available from behavior coding and some of the likely reasons for the problems that the coding revealed. For some of the examples the source of problems was clear simply from examining the question, but for others additional information was needed to diagnose problem sources.

The individual questions described above illustrate some of the problems we faced in interpreting behavior code results. Examination of coding results for all 60 questions led to some general observations about the information provided by behavior coding. In general, for questions with high levels of interruptions or qualified answers, we found that sources of problems identified by behavior coding usually were clear simply from examining the question. As in the first two examples, questions subject to interruption shared a common structural pattern in which an answerable questions is posed and then either additional information or response choices are given. Not surprisingly, this structure appears to encourage respondents to interrupt with their answers after the initial question is posed, or once they hear a response choice they consider suitable.

Questions with high levels of qualified answers were of two types. Some required respondents to report precise information about past events (e.g., Example 5), which placed great demands on their knowledge and recall, and respondents often indicated that their answers were approximate or that they were not sure of their answers. Other questions (e.g., Example 6) required information to be integrated or evaluated -- also demanding tasks.

As the examples show, the information from behavior coding often was insufficient to diagnose sources of difficulties for questions with high levels of requests for clarification, inadequate answers, or "don't know" answers. While requests for clarification usually reflect comprehension difficulty, they do not identify what concept or feature of the question is involved. Additional, more specific, information is needed. Inadequate and "don't know" answers can signal comprehension difficulties or difficulties with performing the response tasks. Again, more specific information than that provided by behavior coding is needed to diagnose the reason(s) for these sorts of answers for a particular question.

Since our goal was to improve questions, we needed additional diagnostic information for a number of the questions. There were several sources of information available to us in this study. One was discussion of behavior coding results with the coders to get their ideas on what was unclear to respondents, in what way responses were inadequate, or why respondents did not know the answer. Members of the research staff familiar with the taped interviews also were consulted. Finally, responses to the special probe questions sometimes helped to interpret behavior coding results. Where needed, we used these sources of information to develop hypotheses about the sources of difficulty for each question with high levels on the problem indicators for Group 1 interviews.

Once probable sources of problems with a question were diagnosed, we revised the question in an attempt to reduce or eliminate the sources of difficulty. The revisions included changes in question structure, clarification of concepts, changes in the response form, and simplification of reporting tasks. Occasionally when a complex concept was involved, the revision replaced the original question with two or more questions, each question in the series covering part of the original concept.

The aim in revision was to maintain the original question objective, replacing the problematic question with one that was easy to understand and that posed a manageable reporting task. Sometimes the objective of a questions was unclear to us as well as to the respondents. In these few cases we used our best judgment as to the objective intended in the survey from which the question had been selected.

Comparison of Respondent Problems with Original and Revised Questions

Table 3 gives an overall picture of the success of the revisions in reducing respondent problems. For purposes of this comparison, we have included in the table only the Group 1 questions that had a 15 percent or higher level on a problem indicator.[4]

[4] The ninth and tenth indicators, which provided no information about the specific types of problem behaviors, were not used at this time to identify problematic questions. Analysis showed that the ninth indicator identified essentially the same questions as problematic as the code-specific indicators. Results for the tenth indicator were less similar to those from the code-specific indicators, primarily because there were a number of questions with

(For three of the original questions the revisions involved splitting the question up into a series of two or more questions; results for these questions are not included in the table and are discussed later.) The table gives results for three questions with high levels of interruptions, 13 questions with high levels of clarification requests, 8 questions with high levels of qualified answers, 33 questions with high levels of inadequate answers, and 3 questions with high levels of "don't know" answers. Some questions had high levels on more than one of the problem indicators.

Table 3 does not include results for the question-asking indicators. As it turned out, most questions with high levels of interviewer reading problems also had high levels of respondent problems. In particular, high levels of major wording changes or discontinuance of question reading parallelled high levels of respondent interruption of question reading. Since these questions were revised to address the respondent problems, interviewer reading problems became mostly irrelevant. With one exception this meant that high levels of question-reading problems did not figure in question revision.

For each respondent problem indicator, Table 3 gives the mean indicator level for the original versions of the questions identified as problematic by the indicator and the mean indicator level for their revised versions. Questions that were found problematic by more than one indicator are represented in the figures for all the appropriate indicators. As the table shows, indicator levels decreased markedly for the revised questions for all indicators but "don't know" answers.[5] Although levels of inadequate answers were greatly reduced, they still had an average of 20 percent for the revisions. This relatively high level may reflect enduring difficulties with recalling and organizing needed information. This issue will be discussed in more detail later.

Table 4 shows the behavior coding results from the second pretest (Group 4) for the revisions of the examples described earlier, and indicates in parentheses the change in results from the first pretest. Examples 1 and 2, about the most recent visit to a doctor and the cost, had high rates of respondent interruptions. In revising these questions we attempted to correct that tendency. The first example was revised 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?" The second example was revised to, "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?" Both revisions were quite successful, with interruptions falling significantly (p<.05) to 8 and 5 percent, respectively.

---

high levels of problems on the specific indicators to which respondents eventually gave adequate answers.

[5] There is a potential problem with interpretation of Table 3 with regard to a regression effect. However, our examination of this issue suggests that it was unlikely to have caused the substantial decline in the problem indicators that we observed.

For Example 3 we hypothesized that the high rates of requests for clarification, inadequate answers, and "don't know" answers stemmed mostly from a lack of clarify of "warning signs or symptoms of cancer." We attempted to describe this concept more clearly in the following revision. "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?" This revision had mixed success. Levels of requests for clarification and "don't know" answers decreased ($p<.20$ and $p<.15$, respectively), but inadequate answers remained at the same level. It appears that either the concept remained somewhat unclear, or the reporting task was too demanding of respondents' knowledge or recall.

Example 4, about care of teeth and dentures, had high levels of inadequate answers. Respondents did not appear to know or remember what form the response was to take, or did not appreciate the importance of responding in terms of the categories. We revised the question to include restatement of the response choices: "How well do you think you do as far as taking care of your teeth or dentures -- very well, fairly well, or not so well?" The revision was successful, with inadequate responses dropping to 13 percent ($p<.05$).

Example 5, asking for the last time the respondent had a general physical examination or checkup, had high levels of inadequate answers. The revision was directed at two hypothesized problem sources. One source of difficulty appeared to be a lack of clarify in the concept "general physical examination or checkup," as revealed by responses to a special probe question. The second source of difficulty appeared to be the lack of a specified response format. To address these difficulties, the question was revised to "The next question is about a general physical examination -- I mean <u>not</u> just to see about some problem or complaint but a general examination. In what month and year did you last have a general physical examination?" The revision was not very successful in reducing problems identified by behavior coding,[6] perhaps because the revision poses the same difficult recall task as the original question. For this and some of the other problematic questions we could not design a revision that was both easy to respond to and met the original question objectives for precise dating. It was therefore not unexpected to find evidence of problems with this revision as well as with revisions of other questions that retain the original difficult reporting tasks.

Example 6, which required respondents to integrate and evaluate considerable amounts of information in order to compare their health to that of others their age, had high levels of qualified answers. For this question it seemed that probabilistic answers, which pose an easier reporting task, would fulfill the question objective. Accordingly, the revision was, "Compared to other people your age, would you say your health is probably better than others, about the same, or probably worse than others." Hardly

[6] Responses to a special probe, however, indicated that respondents had better understanding of the concept.

any respondents indicated uncertainty in response to this question (p<.05 for the change).

Example 7 had high levels of requests for clarification and of inadequate answers. Revision aimed at clarifying what kinds of treatments or examinations respondents were to report. It also specified the required response format, which was missing from the original version. One other source of difficulty was identified by the coders and staff members from listening to the tape recordings: the position of the question in the questionnaire created confusion. It followed a question asking 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 confuse respondents, perhaps making them wonder if they had understood correctly. The question seemed especially confusing for respondents reporting visits to the preceding question.

The revision, designed to address all these issues, included two alternative questions. Respondents who had reported visits to the preceding questions were asked, "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 reporting no visits to the preceding question were asked, "About how many years ago was the last time you were treated or examined for dental care?" By this device, respondents were given a manageably small number of response choices. The revision also clarified that the treatments and examinations were to be for dental care. Levels of requests for clarification and inadequate answers were considerably reduced (p<.05) for these questions.

The questions about sources of heart disease in Example 8 had high levels of inadequate answers. The questions appeared to have poorly designed response categories. The coders indicated that answers were inadequate primarily because respondents merely said "definitely" or "probably," which did not serve to single out one of the response choices. With revisions of these questions that used the response choices "large effect, some effect, little effect, or no effect," levels of inadequate answers dropped considerably (p<.05).

Some of the revisions generated high levels of problem behaviors not evident for the original versions. While there were marked decreases in levels for the targeted indicators, levels for other indicators increased for some of the questions.

Example 9, about servings of butter on a typical day, had high levels of requests for clarification. While respondents may have been confused about what food the servings referred to (butter was the topic of the preceding question and was not restated in Example 9), discussion with coders indicated that respondents appeared to be unclear about the meaning of "a typical day." The revision, "On days when you eat butter, how many servings do you usually have?" apparently clarified the term, with only 4 percent of the respondents asking for clarification (p<.15). However, 17 percent of the respondents gave inadequate responses to the revision, whereas only 7 percent gave them to the original version. We have no explanation for this result.

For most of the adversely affected questions, levels on the affected indicators for the revised versions were only around 15 percent. For several of the questions for which the affected indicator was for respondent interruptions with answers, percentages for the revised versions were somewhat higher, ranging from 15 to 27 percent. With regard to the revisions with newly high levels of respondent interruptions, the revisions shared the same pattern that had encouraged interruption of other questions in the original questionnaire.

Series revisions of three questions

Three of the questions diagnosed as including complex concepts were replaced with question series. These questions were not included in Table 3 because a single score on each index could not be calculated for multi-question revisions. For these questions it seemed particularly difficult to design single question revisions that communicated the original concept clearly. In the revision, each question in the series covered part of the original concept. Problems with one of the three original questions appeared to be mainly conceptual, while with the other two the reporting task also seemed difficult. Behavior-coding results indicated considerable success in problem reduction for the first question. The other two questions continued to show high levels of problems in their revisions, perhaps because the revisions were not designed to simplify the reporting tasks. For one of the questions a high level of respondent interruptions was reduced successfully by the revision. Overall, the strategy for dealing with complex concepts by breaking them into simpler components is promising. One would still expect problems, however, if a difficult reporting task is also involved.

ANALYSIS OF RESPONSES TO SPECIAL PROBES

Introduction

Special probes ask respondents to report their experiences with the questions. Respondents can be asked about the meaning of particular questions, how they went about answering them, and about problems they had in understanding or answering the questions. Some special probe questions can be incorporated in the pretest questionnaire to investigate respondent problems with certain questions; others can be asked when the main interview has been completed.

The major drawback to using special probes in standard pretests is that only a few questions can be probed without unduly lengthening the interview. Also, few questions can be probed immediately following responses without the danger of influencing responses to subsequent questions; moreover, only certain types of probe questions are suitable to be embedded in the questionnaire. Special probes can be included at the end of the questionnaire, without concern about influencing questionnaire responses, although here problems of retrospection enter.

This section describes the use of a range of special probe questions to identify and diagnose problems with survey questions. The original plan was to probe the original and revised questions with the same probes in the two pretests in order that the probe results could help evaluate the success of the revisions. However, after the first pretest it appeared that a number of the probes used were not effective. For the second pretest it seemed best to abandon unproductive probes and to experiment with others that might have greater promise. Accordingly, results for the two pretests are discussed together, with the focus on identifying useful types of probes.

Method

Questionnaire Forms and Special Probes

Special probes were added to the questionnaires for Groups 2 and 4. In order to probe a sizeable number of questions, three forms of the questionnaire were created and about a third of the respondents received each form. This meant that around 33 respondents received a particular probe. In all, somewhat more than a third of the questions were probed in each group.

Each questionnaire form included special probes to be asked immediately following responses. These embedded 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. To avoid disturbing the interview, it was thought best to avoid probes near the beginning and to scatter them throughout the questionnaire. Some of the probes resemble those routinely used by interviewers (e.g., "Could you tell me more about that?"). We judged that none of the probes was sufficiently unusual, nor were they used sufficiently frequently, to disturb the course of the interview.

In each form, 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 an informant rather than a respondent. In the informant role, respondents were asked to talk about their interpretation of the question and report their experiences and difficulties in answering them. This detailed probing focused primarily on the respondent's understanding of question meaning. In addition, probes were designed to ascertain how accurate respondents thought some of their answers were.

We devised probes aimed at three kinds of problems: comprehension of the question, information retrieval, and response category selection. In addition, a number of probes were used that were more general in nature -- that is, not targeted to any particular type of problem. Each form of the questionnaire included a variety of probes, used with a variety of question types.

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 sure that they understand a question but in fact misinterpret it. Question comprehension problems were probed in several ways. One way was to investigate the meanings of particular concepts. These probes were designed specifically for questions. For example, for a question asking about consumption of "red meat, such as beef, pork, lamb, liver, and so on," a probe was used to learn whether the respondent's concept of red meat matched the researchers'.

> Probe: Would you include things like bacon, hot dogs, or lunch meats as red meat?

The second type of comprehension probe asked respondents to elaborate on particular aspects of their answers. For example, for a question asking for the last time the respondent had a general physical examination or checkup, this probe was used to explore the respondent's understanding of "general physical examination or checkup."

> Probe: What was the main reason you went for that visit?

The third type of comprehension probe asked how clear a particular concept was, or how much difficulty the respondent had in understanding the concept. One probe used with a question asking about days that illness "kept you in bed for more than half of the day," was:

> Probe: How clear was it to you what to include as a half day in bed?

Information Retrieval Probes

Some probes were used to reveal difficulties with information retrieval. These asked respondents to talk about how they arrived at their answer, to report problems they had in answering or how hard it was for them to answer, or asked them to assess the accuracy of their answers. For example, for a question asking how long it had been since the respondent had last been treated or examined for dental care, the retrieval process was probed with:

> Probe: How did you figure out when that was?

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?

General Probes

These probes were designed to be general stimuli for additional information that might reveal question problems. These probes were variations on:

> Probe: Could you tell me more about that?

Evaluation of Responses to Special Probes

Calculations were made for each question giving the percentage of interviews in which the probes[7] indicated a problem. Table 5 shows the number of questions probed, the type of probe used and the number of questions with specified levels of problems. Figures are presented separately for the four main types of probes.

Comprehension Probes

Two-thirds of the 18 questions probed for meaning had comprehension problems in 15 percent or more of the interviews. Comprehension probes clearly are capable of revealing problems of understanding and, further, indicate that a sizeable number of questions in the questionnaire were misunderstood by many respondents. The probes revealed misinterpretations of key terms in the question, but did not reveal uncertainty or confusion about question meaning. Respondents did not appear to doubt their own, often mistaken, interpretations.

All three types of comprehension probes revealed this lack of common understanding. 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

[7] For some questions several probes were used with the same respondent.

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 flawed understanding of question meaning.

Another example of a lack of common understanding was provided by responses to the question:

> "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."

This question also was probed intensively. Respondents were asked to identify from a list which health professionals they would include as doctors or assistants. The list 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 disagreement about whether medical advice obtained on the telephone should be included as instances of having "seen or talked to a doctor or assistant about your health." About a third of the respondents thought such contacts should be included, and 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, beginning 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 during the interview were asked:

> 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 <u>at all</u> during that time?"

About a third of the respondents 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.

Comparison of results from comprehension probes to those from behavior coding. Table 6 shows the frequency with which comprehension probes identified problematic questions that were not identified by behavior coding. Twelve questions were identified by the special probes as having levels of comprehension difficulties of 15 percent or more of respondents. Levels of requests for clarification or question repetition available from the behavior coding identified only five of these questions as causing comprehension problems. If respondents were being diligent in their responding role when they did not understand, they should have asked for clarification or indicated their difficulty. The low levels of requests for clarification for most of the 12 questions is further evidence that respondents largely were confident (but often incorrect) as to question meaning. This finding supports the conclusion that the particular strength of comprehension probes is to reveal misinterpretation of question meaning.

Information Retrieval Probes

Fifteen questions were probed for difficulties with recalling and organizing information. Table 5 shows that although the probes provided evidence of recall problems for several questions, only one of the questions appeared to have significant levels of problems. 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 questions had high levels of behaviors often associated with retrieval problems -- inadequate, qualified, or "don't know" answers (see Table 6). 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.

Response Category Selection Probes

Six closed questions were probed for respondent difficulties with selecting the appropriate response category. While respondents might retrieve the information needed to answer a closed question, they might have difficulty mapping that information into the designated response choices. 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.

General Probes

For twelve questions probes were designed to provide a general stimulus for additional information. For two of the questions the probes indicated significant levels of comprehension problems. 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. The probe was:

Probe: Could you tell me more about that?

Responses to the probe indicated that many respondents misinterpreted 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. Again, the probe was:

Probe: Could you tell me more about that?

The responses 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 these results. It may be that these probes are too non-specific and are not sufficiently directed toward potential problem sources. An illustration is provided by another question. Following a question about use of butter, it read:

"What is the number of servings on a typical day?"

The probe was:

Probe: "Could you tell me more about that?"

Responses to the probe described in what ways butter was used.

While more specific probes are more likely to be effective, sometimes the original question provides an adequate frame of reference for the 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 are more useful for revealing comprehension problems than other problems.

CONCLUSION

Analysis of coded interview behavior is a useful technique for obtaining objective information to evaluate questions. The method offers a systematic way of learning the strengths and weaknesses of each question in the questionnaire. In this study behavior coding identified problematic questions and helped in diagnosing the source of the problem. When problematic questions were reworded, coding behavior for the new questions showed decreases in problems.

One weakness of the method is in gaining information on the sources of the respondent problems. For some questions, examining coding results and studying the question itself are adequate to both diagnose problems and suggest solutions. This was true particularly for questions with high levels of interruptions or qualified answers. For other questions, including those with high levels of requests for clarification or of inadequate or "don't know" answers, one may need to rely on information from people who have had experience with the interview.

Special probes are a useful supplement to the behavior coding technique. They can help to illuminate reasons for the problems revealed by coding results. However, the special strength of special probes lies in their ability to reveal problems that are not evident in interview behavior. In particular, special probes can be effective in revealing lack of common understanding and misinterpretation of question meaning.

Since only a limited number of probes can be used in an interview, the special probe technique must be applied selectively in pretests. The investigator needs to decide which questions are particularly important, or which are most likely to be misunderstood, and target them for special probing. Further, the investigator is likely to need to have some idea as to the nature of likely problems in order to probe appropriately. Successful use of the technique depends on the skill of the investigator in identifying questions that merit probing and on skill in designing effective probes.

Revision of questions must still be based on the skill and experience of the person wording the questions. Some questions, however, include complex or fuzzy concepts that defy simplification or clarification. Other questions involve 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 solution is to revise the statement of data required. The investigator may have to give up or substantially revise the aspiration for certain data simply because of the impossibility of the task or the lack of clarity of the concept. In this study, "HMO" is an example of 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. An example of an overly difficult task would be reporting the number of doctor visits over the past five years.

This is the first study to demonstrate that behavior coding and special probes are useful techniques for identifying problems with questions in survey pretests. While the detailed behavior coding system we employed generally worked well, the measures of problem-indicating behavior we derived from it can be achieved with a simplified coding scheme better suited for use in regular pretests. The measures of problem-indicating behavior available from coding provide objective, systematic bases for evaluating questions. Special probes are very useful in learning the sources of respondent problems. In this study we had limited success in devising adequate probes. Clearly, additional work and experimentation should be devoted to techniques and procedures for designing these probes.

REFERENCES

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

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

Cannell, Charles F., Sally A. Lawson, and Doris L. Hausser
A technique for Evaluating Interviewer Performance. Ann Arbor, MI: Survey Research Center, The University of Michigan, 1975.

Cantril, Hadley and Edrita Fried
The meaning of questions. In Hadley Cantril (ed.), Gauging Public Opinion. Princeton: Princeton University Press, 1944.

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

Marquis, Kent H.
Interviewer-Respondent Interaction in a Household Interview. Paper presented at Annual Meeting of American Statistical Assoc., August 19, 1969, New York City.

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

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

Morton-Williams, Jean
The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire. <u>Quality and Quantity</u>, 13, 1979:59-75.

Morton-Williams, Jean, and Wendy Sykes
The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Question. Journal of the Market Research Society, 26, 1984:109-127.

N.C.H.S. (National Center for Health Statistics)
Linked Telephone Surveys: A Test of Methodology. Vital and Health Statistics, Series 2, No. 109, Dept. of Health and Human Services, forthcoming.

Schuman, Howard
The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, 31, 1966:218-222.

Figure 1. Study Design

	Technique(s) used	
	Behavior coding	Special probes
<u>First pretest</u> (with original questionnaire)		
Group 1 (60 respondents, 6 interviewers)	X	
Group 2 (104 respondents 9 interviewers)		X
<u>Second pretest</u> (with revised questionnaire)		
Group 4 (100 respondents 8 interviewers)	X*	X

*A sample of 60 Group 4 interviews were behavior coded.

Figure 2. Behavior Code Categories

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.
B	Break off*,#	Interviewer does not complete reading the question because the respondent has interrupted.

Respondent Behavior Codes

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

*Indicates a problem with the question.

#These two code categories, M and B, were combined and treated as one category for the indices of question problems.

Table 1. Mean Levels and Distributions of Problem Indicators for Original Versions of 60 Questions

Problem indicator	Mean level over the 60 questions*	Distribution of problem indicators#						
		0-4%	5-9%	10-14%	15-19%	20-24%	25-34%	35%+
Interviewer question-reading behavior:								
1. Slight changes	12%	18	12	12	5	3	7	3
2. Major changes	4%	41	13	3	3	0	0	0
Respondent behavior:								
3. Interruption	4%	47	6	2	2	1	1	1
4. Clarification	10%	20	10	15	9	0	6	0
5. Qualified ans.	7%	37	11	3	3	1	3	2
6. Inadequate ans.	24%	4	13	7	10	5	6	15
7. "Don't know"	4%	45	8	4	1	0	1	1
8. Refusal	0%	60	0	0	0	0	0	0
9. Any problem	40%	0	1	3	7	6	11	32
10. No adequate ans.	20%	8	12	11	7	6	5	11

NOTE: The table is based on 60 interviews from Group 1.

*Entries are the percent of times the problem indicator was assigned over all 60 questions.

#Entries are the number of questions (out of 60) with problem indicator scores in the specified ranges of percentages.

Table 2. Problem Indicator Levels for a Selection of Questions

Problem indicator	Question								
	1	2	3	4	5	6	7	8b	9
Interviewer Question-reading Behavior:									
1. Slight changes	8	30	7	2	3	8	5	0	10
2. Major changes	19	17	0	0	2	0	8	2	2
Respondent Behavior:									
3. Interruption	35	23	0	0	0	0	0	0	0
4. Clarification	3	10	27	12	3	3	30	10	17
5. Qualified answer	3	3	12	2	27	22	3	0	2
6. Inadequate answer	8	17	18	28	87	13	30	77	7
7. "Don't know"	0	8	40	2	12	3	3	5	0
8. Refusal	0	2	0	0	0	0	0	0	0
9. Any problem	48	48	70	38	88	33	50	80	20
10. No adequate answer	30	23	20	12	60	25	23	52	2

NOTE: The table is based on 60 interviews from Group 1. Questions are identified by the example numbers in the text.

Table 3. Mean Levels of Respondent Problem Indicators for Problematic Original Questions and their Revisions

		Mean problem indicator level	
Respondent problem indicator	Number of questions	Original Question*	Revised Question#
Interruption	3	28	4
Clarification	13	22	11
Qualified answer	8	27	9
Inadequate answer	33	37	20
"Don't know"	3	30	31

NOTE: The three questions for which revisions were series of questions are not included in this table because of the difficulty of calculating one score for each indicator for the revisions. Each row is based on questions for which 15 percent or more respondents exhibited the indicated behavior for the original question. A question may figure in more than one row.

*Group 1 interviews.
#Group 4 interviews.

Table 4. Problem Indicator Levels and the Amount of Change for Revised Versions of Selected Questions

	Question								
Problem Indicator	1	2	3	4	5	6	7	8	9
Interviewer Question-Reading Behavior:									
1. Slight changes	8(0)	13(-17)*	10(+3)	2(0)	17(+14)*	18(+10)	10(+5)	13(+13)*	9(-1)
2. Major changes	10(-9)	12(-5)	3(+3)	15(+15)*	3(+1)	0(0)	15(+7)	27(+25)*	0(-2)
Respondent Behavior:									
3. Interruption	8(-27)*	5(-18)*	0(0)	5(+5)	2(+2)	0(0)	8(+8)*	12(+12)*	0(0)
4. Clarification	7(+4)	18(+8)	17(-10)	2(-10)*	13(+10)*	3(0)	6(-24)*	8(-2)	4(-13)
5. Qualified answer	0(-3)	2(-1)	7(-5)	0(-2)	23(-4)	2(-20)*	2(-1)	3(+3)	0(-2)
6. Inadequate answer	17(+9)	18(+1)	18(0)	13(-15)*	77(-10)	5(-8)	15(-15)*	12(-65)*	17(+10)
7. "Don't know"	0(0)	15(+7)	27(-13)	0(-2)	15(+3)	7(+4)	0(-3)	3(-2)	4(+4)
8. Refusal	0(0)	0(-2)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
9. Any problem	29(-20)*	43(-5)	43(-27)*	20(-18)*	82(-6)	15(-18)*	29(-21)*	30(-50)*	22(+2)
10 No adequate answer	10(-20)*	18(-5)	10(-10)	10(-2)	57(-3)	3(-22)*	15(-8)	18(-34)*	0(-2)

NOTE: Questions are identified by the example numbers in the text.
The first number in each cell is the problem indicator level for the revised version of the question (Group 4). The number in parentheses gives the amount and direction of change from the original version of the question in the first pretest (Group 1).

* p<.05, two-tailed test.

Table 5. Number of Questions Probed with each Probe Type and Evidence of Problems in Responses to the Probes

Probe type	Total number of questions probed	Level of evidence of problems*		
		≥15%	10-14%	≤9%
Comprehension	18**	12	1**	5
Information retrieval	15	1	8	6
Response category selection	6	1#	3#	2
General	12	2	0	10

NOTE: Results of both pretests are included, with original and revised versions of questions treated as separate questions.

* Levels are percentages of respondents with evidence of the type of problem probed.

**Seven questions in the original questionnaire that were simultaneously probed with one special probe directed at understanding of a shared concept (heart disease) are treated as one question in this analysis.

Problems revealed for these questions were with recalling information, not with category selection.

Table 6. Comparison of Identification of Problematic Questions by Special Probes and by Behavior Coding

	Results from behavior coding*	
	Not problematic	Problematic
Results from comprehension probes:		
Not problematic	6	0
Problematic	7	5
Results from information retrieval probes:		
Not problematic	5	9
Problematic	0	1
Results from response category selection probes:		
Not problematic	1	4
Problematic	0	1**
Results from general probes:		
Not problematic	6	4
Problematic	0	2

*Entries are numbers of questions.

**Problems revealed by the probes were with recalling information, not category selection.

CHAPTER 4

Evaluation of Special Training and Debriefing Procedures for Pretest Interviews

Floyd J. Fowler, Jr.

INTRODUCTION

In a typical pretest of a survey instrument, experienced interviewers carry out 20 or 30 interviews with people similar to those to be included in the survey. The interviewers then participate in a debriefing session in which they report to the investigators any "problems" they encountered in doing the interview. There are a number of potential limits to such a procedure:

- The criteria for what kinds of problems interviewers are supposed to identify usually are not well defined.

- Interviewers typically receive little or no specific training in how to pretest interview schedules and how to identify experiences indicative of problem questions.

- The debriefing process itself seems likely to be an imperfect way of identifying problems, with the views of some interviewers being likely to be more prominent or frequently expressed in the debriefing than others because of their personal style rather than the quality of their observations.

A more thorough discussion of the weaknesses of traditional pretest procedures is presented in Chapter 2.

There are at least two consequences of inadequate pretesting procedures. On the one hand, there is reason to be skeptical that a high proportion of problem questions are in fact identified through this process. Second, because of the qualitative and subjective process for problem identification, it is easy for researchers to ignore or misinterpret the input they get from a pretest.

The main emphasis of this research was to develop and evalute methods that do not rely on reports by interviewers. One of the methods is based on systematic coding of interviewer and respondent behavior in pretest interviews, presented in detail in Chapter 3. However, the traditional pretesting procedures have a well established place in the survey research process. Hence, we thought it would be valuable to see if the traditional interviewer-oriented pretest could be improved to address some of the limitations noted above. To this end, another component of our research included specially developed, and we hoped improved, procedures for training and debriefing pretest interviewers. In this chapter, we report this research and show how question problems identified by special procedures compared with those identified in a traditional pretest and also how they compared with the problems indicated by systematic coding of interview behavior.

METHOD

The special treatment of pretest interviewers (Group 3) consisted of two main parts: specific training in pretest procedures and amended debriefing procedures that included completion of a standardized question rating form.

The study involved five interviewers randomly selected from the telephone interviewing staff from the Center for Survey Research at the University of Massachusetts-Boston. They included four women and one man, ranging in interviewing experience from less than one year to almost 20 years.

A special interviewer training program lasted about five hours. During the first hour, the role of the pretest interviewer was discussed, the problem with the limited information usually obtained from a pretest was presented, and the desire to make interviewer pretesting more systematic was emphasized.

Four kinds of problem questions were discussed in detail:

- Questions that interviewers have difficulty reading as written, or think respondents have trouble understanding as written.

- Questions that include terms or concepts that are not understood consistently by all respondents.

- Questions that pose problems for respondents in knowing what the answer is, either because they are asked for information that is difficult to recall or asked for an opinion or attitude about something that they have not thought about before.

- Questions that pose difficulties because of the way they are to be answered. Examples included questions with response alternatives that do not fit the question; questions that do not specify what kind of answer will suffice; and questions that do not specify the level of precision that is required of the answer.

After discussing these issues in a general way, the balance of the training consisted of listening to tape-recorded interviews. As a group, interviewers discussed what they heard and tried to agree on the kinds of problems that interviewers and respondents had with the questions. A specific goal was to get interviewers to be attentive to whether or not interviewers were reading questions as written and to think about how question wording contributed to the various problems that respondents had in answering questions.

Once interviewers had been trained, they were given an assignment of telephone numbers drawn from directory listings in Southeastern Michigan and asked to complete ten interviews. They were to make a maximum of two calls to find people who were not at home. They were given a procedure for choosing among household members at home, to increase the percentage of respondents who were male and who were young. With respondent permission, interviewers tape recorded all interviews so that the interviewer and respondent behavior could be coded from the recordings.

After each pretest interview, interviewers were instructed to go through the interview schedule and to make notes on a master interview schedule of places where they or the respondent had difficulty. Once all their ten interviews had been completed, interviewers were instructed to complete a standardized rating form (Appendix D). On this form, they were asked to rate each question with respect to each of the four types of problems described above:

A. No problem evident.

B. Possible problem evident.

C. Definite problem evident.

The debriefing session was led by an experienced interviewer supervisor who directed the interviewers in discussing each question. Although the discussion was similar to usual debriefing sessions, because interviewers presented their ratings as part of the discussion, the debriefing session took considerably longer than usual. For this 20 minute interview, the debriefing session lasted about two and a half hours. After the debriefing, interviewers were asked to complete the rating form again based on any additional thoughts or insights they had gathered during the course of the discussion.

Interviewer ratings of each of the four types of problems were summarized for each question to provide an overall rating for each problem type. Pre- and post-ratings were summarized separately as follows:

<u>No problem</u>. At least four of the five interviewer ratings were A (no problem).

<u>Definite problem</u>. At least two interviewers' ratings were C (definite problem).

<u>Possible problem</u>. All other patterns of interviewer ratings.

The tape recordings of the 50 interviews from this pretest were coded by a staff of trained coders at the Survey Research Center (SRC), The University of Michigan (see Chapter 3). Although a number of aspects of behavior were coded, three key measures will be the focus of this analysis:

1) Whether the interviewer read the question exactly as worded, with only minor changes that did not affect the question meaning, or made major changes in question wording when reading the question,

2) Whether or not the respondent asked for clarification,

3) Whether or not the respondent gave an answer that did not meet the question objectives (an inadequate answer).

Occurrences of each of these behaviors were coded, and rates were calculated as the number of interviews in which the particular behavior occurred divided by the number of interviews in which the question was asked. One focus of the analysis will be the way that the specially trained pretest interviewers' ratings compared with the results of this behavior coding.

The other comparison in this paper will be with the results of a pretest by interviewers at the Survey Research Center at The University of Michigan using the same questionnaire (Group 1). Interviewers at Michigan interviewed respondents drawn from the same sample frame, using the same procedures for respondent selection.

Michigan interviewers received standard SRC instruction in pretest procedures, and participated in a standard debriefing session. This included a question-by-question discussion of question problems. There was no specific, systematic written evaluation of questions, though interviewers did record comments about questions in "comment copies" that were available for review.

In this analysis, the key source of data about the Michigan pretest experience comes from content analysis of a tape recording of the debriefing session. For each question, a rater made two kinds of determinations (Appendix B):

1) Did the interviewers mention a problem with reading the question?

2) Did interviewers report that respondents had difficulty understanding the question or how to answer it?

The criterion for classifying something as a "problem" was either that two or more interviewers mentioned it, or that one interviewer said that a problem was apparent in at least two interviews.

The analyses that follow are designed to address two main questions:

1) How do ratings of questions by specially trained pretest interviewers (the Boston interviewers) compare with results of the traditional pretest experience (Michigan interviewers)?

2) How do ratings of questions by specially trained pretest interviewers compare with the results of behavior coding to detect question problems?

RESULTS

Table 1 compares the extent to which the specially trained interviewers and the interviewers who received no special training identified problems with questions. For specially trained interviewers, problem identification was based on their ratings of question problems. Problem identification for the regular interviewers was based on the coding of their debriefing.

The first part of the table compares the findings with respect to question reading. It is clear that the ratings by specially trained interviewers identified more questions as being hard to read than the Michigan interviewers mentioned in their debriefing. There are several possible explanations for this difference. One explanation is that the special training sensitized interviewers to reading problems. However, when a content analysis of the Boston debriefing was done using the same coding rules that were used for the regular debriefing, it was found that Boston interviewers reported <u>fewer</u> questions as having reading problems than the regular interviewers (9 vs. 10 questions). It seems likely, therefore, that some feature of the question rating method used by the Boston interviewers accounts for the increased number of problem questions identified by that method, rather than the special training the interviewers received. In particular, the description of "difficult to read as written" on which Boston interviewers based their ratings seems more inclusive than the definition on which analysis of the debriefing was based.

The last part of Table 1 compares findings about respondent problems. For this comparison a single measure of respondent problems was constructed by combining the three specific types of problems rated by the Boston interviewers. This produced a measure more comparable to the measure of respondent problems available from the regular debriefing. The results are very comparable. All of the questions identified as definite problems by the specially trained interviewers were also identified as problems by the regular pretest interviewers; of the 23 questions that were labelled as "no problem" by the specially trained interviewers, the regular pretest interviewers agreed on 18 of them. If one treats a "possible" problem as a question that has been flagged, the two groups of interviewers disagreed about only 9 of the 60 questions.

Table 2 addresses the question of the extent to which the behavior coding and the interviewers identified the same or different questions as problems. The table shows considerable correspondence between results based on ratings by the specially trained interviewers made before their debriefing and behavior coding results. However, a number of questions were identified as problematic by one of the methods that were not identified by the other, and Table 2 makes us clearly address the issue of "cut points". Interviewers, of course, have to decide when a question is troublesome enough to flag it as a "problem." We adopted a rule that if two interviewers out of five rated a question as having a problem, there was a problem; other rates obviously are possible. In the same way, the behavior coding yields a continuum of rates of various behaviors; researchers have to decide when a particular rate constitutes a "problem" for a particular question. These issues are highlighted in Table 2 because the main difference between the specially-trained interviewer ratings given before the debriefing discussion and after was that they decided that more questions should be labeled "problems".

Looking just at the pre-debriefing ratings of the specially trained interviewers, on average they flagged about two-thirds of those questions that the behavior coding indicated to be problems (using the 15 percent cutoff for a "problem"). At the same time, they flagged about a fifth of the questions as problems that did not meet the 15 percent standard for behavior. If one uses the ratings after the debriefing discussion, one gets a much higher percentage of the questions "flagged" by the behavior coding rated as problems by the interviewers. However, when interviewers increase their level of sensitivity, they also increased the number of the questions that did not meet behavior coding standards as being possible problems as well.

DISCUSSION

One obvious conclusion from these data is that interviewers are not bad observers of interview problems. They obviously are identifying and reporting on many of the same kinds of problems that one derives from behavior coding.

Another possibly surprising result of these analyses is that the special training did not markedly affect identification of problems with questions. The more general question addressed in this research is how best to use interviewers in the pretest process. With respect to the value of interviewer training, these data suggest that the three or four hours invested in trying to sensitize interviewers to question problems were not particularly productive.

The hope was that interviewers would be sensitive to the different kinds of problems that questions posed for people. The specially-trained inteviewers were supposed to specifically focus on problems with understanding the concepts, problems with obtaining the information to answer the question, and problems with the response task itself. However, the interviewers had a hard time making these distinctions. The analyses here do not enable us to assess whether the specially-trained interviewers were more sensitive and refined in their problem identification. Our guess is that they were not.

The other feature of the special pretest was that interviewers filled out standardized rating forms for questions. We do think rating forms are a good idea. While correspondence between what came out of the regular pretest debriefing and the results of the ratings are encouraging in that standard debriefing processes may identify most of the problems that interviewers have to report, the ease of using systematic ratings for tabulating problems has much to commend it. Such forms involve virtually no cost. In the future, we would amend the rating form to have only three ratings: difficulty in reading the question as worded, whether the question contains unclear or poorly defined terms, and respondent problems in answering the question or in performing a response task.

An interesting question is whether or not interviewers should rate questions before the debriefing discussion, after the debriefing discussion, or both. Although the results are highly correlated, of course, more questions are rated as problems after the discussion than before. Interviewers typically take only a few pretest interviews, in this case ten. When a respondent or two has a problem with a question, interviewers have to make a judgment about whether that was simply an idiosyncratic response of those respondents or whether there really is something about the question that will pose a similar problem for a significant number of respondents. When interviewers hear that other interviewers had similar experiences, it increases their confidence that their own experience was not unique and is something that researchers should tend to. However, there was some tendency for correlations with behavior codings to be a little lower based on ratings after the discussion as compared with before which indicates that, as one might expect, there may be some error introduced by the group discussion process.

At this point, our recommendation would be to have interviewers rate questions before and after discussion. Having the ratings made before the discussion serves to systematize and organize interviewers' thoughts, and forms a good basis for the discussion of questions. Ratings made after the discussion allow interviewers to take a wider range of experience into account. Investigators could use both sets of ratings in question evaluation.

Finally, we return to the question of what behavior coding adds to interviewers and what interviewers add to behavior coding. Although clearly the interviewer ratings and the results of behavior coding are correlated, there is no question that behavior coding picks up some kinds of problems that interviewers miss and that interviewers may identify problems that behavior coding misses. In addition, there is the undeniable fact that behavior coding is much more systematic and replicable than interviewer reports and is independent of the perceptiveness of the individual interviewers who are doing a pretest. From the point of view of a researcher, it seems almost certain that he or she would have more confidence in what had come out of the pretest experience if he or she had access to the results of behavior coding.

In general, we conclude that interviewer ratings and behavior coding constitute two different methods of identifying question problems. We believe the best procedure is to have ratings from interviewers and results from behavior coding. In that way, the researcher has two opportunities, rather than just one, to make sure that a question that needs attention is not ignored.

TABLE 1. Comparison of Number of Questions with Problems Identified by Specially Trained Interviewers and by Regular Pretest Interviewers

Problem rating by specially trained interviewers*	Problem identified in regular debriefing	
	Yes	No
<u>Difficult to read</u>		
Definite problem	2	18
Possible problem	2	9
No problem	6	23
<u>Any problem answering</u>		
Definite problem	14	0
Probable problem	19	6
No problem	3	18

*These were interviewer ratings after the debriefing discussion.

Table 2. Comparison of Problem Questions Identified by Specially Trained Interviewers and by Behavior Coding

Rates from coding interview behavior	N	Percentage of questions identified as problems*	
		Pre-debriefing rating	Post-debriefing rating
Read questions exactly			
>.75	15	20%	34%
.76-.85	10	50	60
<.76	35	80	73
R asks for clarification			
<.10	37	14	32
.11-.15	9	33	33
>.15	14	64	79
Inadequate answers			
<.15	26	27	38
>.16-.25	15	53	87
>.25	19	47	79

*Percentage of questions rated a "possible" or "definite" problem.
Figures in the section of the table labeled "Read question exactly" are based on ratings of whether the questions were "difficult to read." Figures in the section labeled "R asks for clarification" are based on ratings of whether the questions contained terms or concepts that were not consistently understood by respondents. Figures in the section labeled "Inadequate answers" are based on an integration of ratings of whether the questions posed difficulty for respondents in knowing the answer and ratings of whether the questions posed difficulty for respondents in providing the answer required.

CHAPTER 5

The Significance of Unclear Questions[1]

Floyd J. Fowler, Jr.

INTRODUCTION

It is axiomatic that survey questions should be clear; they should mean the same thing to all respondents. In practice, however, it is not easy for researchers to know when their questions include unclear terms.

The behavior coding scheme described in Chapter 3 focuses on the behaviors of the interviewers and the respondents. Each question produces an interaction which we think of as a set of turns: the interviewer asks a question, the respondent says something, then the interviewer says something, then the respondent says something until the respondent finally gives an adequate answer or the interviewer gives up. The ideal question would always be read exactly as written and answered adequately on the first try by the respondent; that would be one interviewer turn and one respondent turn. Deviations from this ideal are likely to be meaningful indications of an imperfect question.

The particular focus of this paper is on the value of such coding to identify unclear concepts. This paper is aimed at demonstrating three points:

1. Behavior coding is a useful way of identifying problems.

2. Questions can be rewritten to clarify terms and reduce those problems.

3. Unclear terms are significant sources of error in estimates based on surveys.

As described previously, the focus of this research was a health interview survey consisting of 60 questions drawn from instruments used in several national health surveys mainly by academic or government organizations. A total of 110 such interviews (Groups 1 and 3) were tape recorded. For each question asked in each interview, trained coders noted whether or not the interviewer read the question exactly as worded or made changes. They then coded the respondent's behaviors after the question was read as described in Chapter 3.

The results of this behavior coding were studied for evidence of problems with questions. One indication of a question problem was whether or not respondents asked for clarification in at least 15 percent of the pretest interviews; another was whether or not at least one inadequate answer was

[1] A paper presenting these analyses was given at the Fifth Conference on Health Survey Research Methods, May 2-5, 1989. Proceedings will be published by NCHSR.

given in at least 15 percent of the interviews. Although the selection of 15 percent was arbitrary, it proved a reasonably easy task to identify ambiguities and problems with questions that met those criteria.

There were several kinds of question problems that led to comprehension difficulty. In a few cases, the problem seemed to be primarily with the order of the words, so that respondents had difficulty retaining all the parts of the question that they needed to remember in order to answer the question. The solution to such problems usually was to reorder the question.

There also were questions in which the reason for requests for clarification and/or inadequate answers appeared to be that at least one key term in the question was ambiguous. In those cases, the solution was to change the question to clarify the meaning of the key terms. This paper is focused on five such questions. It should be noted that we did not change the basic objectives; hence questions that posed a task for respondents that they could not perform easily would not be improved by these changes.

The revised survey instrument was readministered by new groups of interviewers using procedures identical to those in the initial phase of the project (Groups 4 and 5). Respondent samples were drawn from the same sample frame. The 150 interviews again were tape recorded and behavior was coded for a subset of 60 Group 4 interviews. The results presented here focus on the effect of the revised wording on the rates at which respondents ask for clarification, the rates at which inadequate answers were given, and the distribution of responses to the questions.

RESULTS

There is not an easy way to discuss the questions in aggregate, because the problems posed by each of the original questions were different. Therefore, we present the issues and results for five of the questions studied, one at a time.

The first question considered deals with the consumption of eggs. After answering a question about how many days each week they had eggs, people were asked how many servings they ate on a typical day. There were requests for clarification in almost a third of the interviews. One main ambiguity lay in what constituted a serving; "typical" may also have been unclear.

When the question was revised to ask people how many eggs they ate on the days when they ate eggs, there were major effects both on the distribution of answers and on interview behavior. It is clear that many, but not all, people thought that a serving of eggs equalled two eggs; others thought it equalled only one. In any case, one gets a very different distribution, and one would guess a more interpretable distribution, of egg consumption with the revised question. At the same time, requests for clarification and inadequate answers drop to zero (Table 1). This appears to be a clear example of how an unclear concept shows up in coding pretest behavior, and how clarifying a term affects the interview behavior and the distribution of answers.

The next question asks about the consumption of butter (Table 2). One ambiguity in that question was whether or not margarine counts as butter. In the original question, there were requests for clarification in over 15 percent of the interviews and inadequate answers were given in 15 percent of the interviews. When the question was revised to specifically exclude margarine, there was a very significant decrease in the number of days that respondents said they had any butter at all.

The changes also may have affected requests for clarification and inadequate answers though the effects were very small at best. This pattern illustrates the fact that unclear concepts are only one cause of these behaviors. A major cause of inadequate answers to this question was that people were supposed to come up with an exact number of days, and in some cases they found that a difficult task. The clarification of what was included in butter did not affect the difficulty of the response task itself.

Table 3 presents a question that pertains to reported exercise, an important focus of health behavior surveys. In fact, the rate of requests for clarification was not high, but the rate of inadequate answers made the 15 percent level. One issue seemed to be what counted as exercise and whether or not walking counted.

The revised question seemed to have an effect on the distribution of answers. The change from 48 to 60 percent who said they exercised regularly is nearly statistically significant. From a behavioral perspective, there also was a reduction both in the rates of requests for clarification and in the rates of inadequate answers.

The question in Table 4 asked whether the last visit to a doctor occurred at a health maintenance organization (HMO). Both requests for clarification and the rate of inadequate answers suggested a problem with the question, and one part of the problem seemed to be understanding what constituted an HMO. The question was revised to clarify that. Also, it was broken into two question, the first pertaining to whether or not respondents belonged to an HMO, the second about whether or not the last visit to a doctor was through the HMO plan.

This change appears to have had a significant effect on the responses, with fewer people reporing their most recent visit was through an HMO. In addition, there was a marked decrease in the rate of requests for clarification of this question, and there probably were fewer inadequate answers, despite the fact that people were now answering two questions, which gave them twice as many opportunities to give inadequate answers.

Finally, Table 5 presents the results for a standard question regarding disability days over the preceding year. For its initial form, 15 percent of the pretest respondents requested clarification. One problem seemed to be ambiguity about what was meant by half a day; there also seemed to be confusion about whether or not extra time in bed for vague maladies (rather than specific conditions) should be counted.

The revised question, designed to clarify those two points, appears to have changed the answers. In particular, there were more people reporting eight or more disability days. The mean was higher, too, though it did not reach the .05 level of significance given the sample size and variance. The fact that the distribution changed is evidence that there is important ambiguity, since the two questions should be equivalent in meaning. However, it is not clear that the new version is a better question. Based on the coding of pretest behavior, it may well be a worse question. Alternatively, the real problem with the question, that it poses a virtually impossible recall task, may be much more apparent to respondents, and show up more clearly in the coding, when the meaning of what is really wanted is clarified.

DISCUSSION

The contribution of this paper is to show the value of behavior coding in identifying question problems that can produce seriously biased estimates. As the data clearly indicate, unclear terms not only produce random error; they can produce systematically distorted results. Moreover, when terms are clarified, the results almost certainly are more accurate (i.e., correspond better with what the researcher is trying to measure) and the improvement can be apparent in the behavior coding as well.

It should be recalled that all the items in this study were drawn from surveys done by established survey organizations. Presumabley, all items had been subjected to standard pretest procedures. Yet, at least 10 percent of the 60 items had a key term or concept that was unclear enough that it met our standard as a "problem" based on behavior coding. Although clarifying the terms did not always change the distribution of answers, it often did.

These issues are particularly relevant for factual questions -- questions that ask about the occurrence or frequency of events or behaviors. "Whatever it means to you" is not an adequate approach to asking about butter consumption or the number of visits to doctors. For such questions, the researcher and all respondents must share a common understanding of what is and is not to be reported. Otherwise, as we can see, some respondents are reporting margarine use, others not, and the researcher has no idea what has been reported. The data are nearly meaningless.

It should be noted that the researcher cannot rely on the interviewer to clarify poorly defined terms. Respondents will not consistently indicate how they interpret a question. Rather, the problem must be identified during pretesting, before the actual survey, so that all respondents are exposed to the same, clear question.

All problems of comprehension do not show up in behavior coding. Sometimes respondents will answer questions they do not understand without asking for clarification. We think focused group discussion and laboratory studies should be done before a survey instrument is subjected to formal pretesting. Intensive reinterviews and "think aloud" interviews are particularly good ways to identify comprehension problems. However, once the developmental work is done and the instrument is ready to be tested in a realistic interview setting, coding behavior is an objective, effective and reasonably low-cost way to identify significant remaining problems with questions.

Table 3. Comparison of Answers and Interview Behaviors for Two Versions of Question on Regular Exercise

Original Q: Do you exercise or play sports regularly?

Revised Q: Do you do any sports or hobbies involving physical activities, or any exercise, including walking, on a regular basis?

		Original Q	Revised Q
		Percentage of answers	
Regular exercise:			
Yes		48	60
No		52	40
		100%	100%
	n	110	150
		Percentage of interviews in which behavior occurred	
Requests for clarification		5%	0%
Inadequate answers		20%	12%
	n	110	60

Table 4. Comparison of Answers and Interview Behaviors for Two Versions of Question on whether Last Doctor Visit was at an HMO

Original Q: Was that place a health maintenance organization or health care plan, that is, a place you go for all or most medical care, which is paid for by a fixed monthly or annual amount?

Revised Q: Do you belong to an HMO or health plan that has a list of people or places you go to, in order for the plan to cover your health care costs?

Was your last visit to a medical doctor covered by your health plan?

		Original Q	Revised Q
		Percentage of answers	
Last see doctor at HMO:			
Yes		39%	23%
No		61	77
		100%	100%
	n	110	150

		Original Q	Revised Q
		Percentage of interviews in which behavior occurred	
Requests for clarification		17%	2%
Inadequate answers		27%	18%*
	n	110	60

*12% and 6% respectively for the two revised questions.

Table 5. Comparison of Answers and Interview Behaviors for Two Versions of Question on Disability Days

Original Q: 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? Include days while an overnight patient in a hospital.

Revised Q: The next question is about extra time you have spent in bed because of illness or injury, including time spent in the hospital. During the past 12 months since July 1, 1987, on about how many days did you spend several extra hours in bed because you were sick, injured, or just not feeling well?

	Original Q	Revised Q
	Percentage of answers	
Number of days:		
0	57	48
1-7	36	33
8 or more	7	19
	100%	100%
Mean number of days	2.6	4.0*
n	110	150
	Percentage of interviews in which behavior occurred	
Requests for clarification	15%	17%
Inadequate answers	7%	30%
n	110	60

*There was one person who reported 90 days in the second sample, almost twice as many as the next person in either sample. Removing that person reduces the mean to 3.4. In either case, the difference is not statistically significant.

CHAPTER 6

Conclusions and Recommendations

INTRODUCTION

The essence of survey research is the collection of information using a standardized questionnaire. Although the questionnaire is the measuring instrument upon which the success of the whole survey operation depends, its development and testing are the least scientifically rigorous component of the survey process. Despite the valuable research on question form and response mode issues conducted by a number of investigators (e.g., Cantril, 1944; Payne, 1951; Schuman and Presser, 1981; Sudman and Bradburn, 1982), the creation of a survey questionnaire remains largely an art, based primarily on past experience with only a few "common sense" principles as guidance.

A survey questionnaire may go through several developmental stages between the researcher's formulation of objectives or hypotheses and data collection. For topics with which the investigator has little experience and those that pose particularly difficult problems for the respondent, substantial developmental work is needed prior to a formal pretest. This preliminary activity provides a basis for understanding the types and levels of questions that can be asked. The use of such techniques as focus groups, depth interviews, and open discussions with individuals or groups will assist the researcher to learn the terminology and language level and the level of understanding of the topics being studied. In recent years organizations have begun to use more formal procedures for the developmental phases of questions and questionnaires. "Cognitive" or "questionnaire development" laboratories may be used for individual or group interviews that may include studies of information storage and retrieval mechanisms as well as exploring respondent comprehension and recall.

Based on the developmental work, a questionnaire is designed and a formal pretest is conducted. The pretest provides the opportunity to test the adequacy of the questions under realistic data collection conditions with representative respondents and interviewers.

While survey researchers agree on the need for pretesting, little attention has been given to pretest methodology. We reviewed some of the most commonly used texts in survey methods, and in each case authors exhort researchers to pretest but give little advice on pretesting procedures.

Chapter 2 describes pretest procedures in common use. The usual strategy is to have interviewers administer the questionnaire to some 25 to 75 respondents using standard interviewing techniques. Experienced interviewers are usually selected for pretesting. They are instructed to be alert to problems that respondents appear to have in answering questions, but otherwise are given no special training or instructions. After the pretest interviewing is completed, the interviewers meet with a supervisor and the survey investigator in a group debriefing session in which the questionnaire is reviewed question by question with interviewers reporting any problems they have observed.

The debriefing is the principal vehicle through which interviewer evaluations are obtained. What transpires during the sessions is critically important to identifying and correcting question problems. However, research data and practical experience with the interaction of group participants lead one to be concerned about group debriefing as a valid procedure for evaluating a questionnaire. For example, even when group members have different opinions they tend to arrive at a consensus, resulting from the influence of dominant members. Opinions of quieter members may never be expressed. The style and status level of the discussion leader also has substantial effect on the nature and quality of the outcome.

In this study (Chapter 2) we compared two group debriefing sessions for the same questionnaire. This comparison demonstrates weaknesses of customary debriefing procedures for producing reliable and useful information. The two groups did not uniformly identify the same questions as problematic. Even when the same questions were identified, the nature of the problems reported varied in the two debriefings. It was difficult to judge the prevalence or seriousness of problems since interviewers did not all report their experience with each question, and their statements of problems were often vague, non-quantitative judgments.

The general conclusion is that the debriefing is unlikely to provide the investigator with an accurate picture of the problems with the questions. Such unreliable, unsystematic evaluations are an inadequate technique for developing a scientific measuring instrument. This situation led us to attempt to devise and test some methods of pretesting that would provide investigators with more objective and systematic information for evaluating questions.

EXPERIMENTAL PROCEDURES

The methods are directed at identifying a variety of problems that frequently lead to invalidity in survey data. In summary, these include --

- Interviewer problems with asking questions as worded.
- Respondent problems of comprehension, either uncertainty as to meaning or a lack of common understanding among respondents.
- Respondent problems with knowing or providing the required answers.

Three techniques for detecting questions with these problems are examined in this research. They are:

1. *Behavior coding.* Aspects of interviewer and respondent behavior in the pretest question-answer process that indicate problems are coded as a means of identifying questions that need to be reworded or redesigned.

2. *Special probes.* Follow-up probes are added to the pretest questionnaire to assess respondents' understandings of questions and specific terms used, and to investigate response difficulties.

3. *Special training of pretest interviewers and question rating.* Pretest interviewers are given training in how to recognize problems with questions and identify the nature of the problems. After completing their interviews, they are asked to rate each question to identify different types of problems they observed.

Each of these techniques is described in detail in preceding chapters: behavior coding and special probes in Chapter 3, and interviewer training and rating of questions in Chapter 4. As these chapters show, each of the techniques is useful for identifying question problems. The three techniques provide different but complementary information. Their combination is an integrated system holds great promise for improving pretesting.

RECOMMENDATIONS

We conclude this report with recommendations for an integrated program for pretesting questions. Some of these suggestions come directly from our research findings while others are not research based but are derived from our study of methods in general use and from our experience in conducting this study.

Our proposed structure for a pretest does not vary greatly from that presently in common use. We propose that the pretest be based on around 50 interviews with respondents similar to those to be selected in the final sample. At the completion of interviewing, an interviewer group debriefing is held. The debriefing is, however, significantly changed from the usual format by the introduction of data from behavior coding and interviewer ratings of questions. These new inputs radically change the process by which question evaluations are carried out in the debriefing session. The roles of the debriefing moderator and the interviewers are also significantly altered. Free-flowing discussion of question problems is replaced by one directed by more objective, quantitative information.

We urge the use of behavior coding and interviewer ratings to identify questions problems. These procedures are simple and inexpensive, and provide objective data on frequencies of types of question problems. Summaries of these data, introduced into the debriefing, can help to organize and focus the discussion. We also recommend that each pretest interviewer record question problems he or she identifies on a "master questionnaire." These questionnaires may then be used in the debriefing to remind interviewers of problems they faced. With information from these sources available, the debriefing is a more effective and efficient procedure. The effectiveness can be maximized by training the debriefing moderator in skills of effective group interactions and decision-making.

We recommend also the use of special probes to assist in understanding the bases for problems. Such probes should be directed principally at comprehension of terms and concepts.

In the following sections we specify the use of these techniques in greater detail.

Behavior Coding Pretest Interviews

One of the main findings of this research is that question problems often have consistent and meaningful effects on the behavior of interviewers and respondents. Coding behavior can be used to identify problems with questions and provide data on the prevalence and nature of the problems. Such coding identifies questions that are not read as written, that are not consistently understood, or that are difficult to answer.

Chapter 3 describes the detailed coding scheme used in this experimental study. The measures of problem-indicating behavior we recommend for regular pretesting, however, can be achieved using a simplified coding scheme. The specific codes that we recommend and the definitions of each category are given in Figure 1. To code behavior for a question, the coder need note only whether or not each type of behavior occurred. Figure 2 shows a simple form on which a number of interviews can be coded simply by making check marks in relevant boxes. Rates of problem-indicating behaviors can be calculated as the number of interviews in which the particular behavior occurred divided by the number of interviews in which the question was asked. The denominator (number of times the question was asked) is obtained by adding the number of checks in the three "Question-asking" boxes. In the indices of question problems used in our research, a major change in question reading included cases in which the interviewer did not complete reading a question because the respondent interrupted with an answer. To avoid overlap in information between the major change and respondent interruption indicators, we recommend that discontinued question readings be coded as major changes only if a major change was made in the portion of the question that was read.

The coding system was developed on the basis of experience with a particular set of questions. The codes were designed to be generally applicable to survey questionnaires. However, the code categories can be easily modified or extended for particular questions and special objectives. For example, some questionnaires include many questions for which the interviewer must fill in a name, date, or some other content that changes according to the immediate interview situation. A question-reading code could be devised to detect errors in selection of the appropriate word or phrase. The system is flexible, and can be adjusted or extended to meet special requirements of particular questionnaires.

Coding can be done live or from tape recordings. Whichever is done, coding with the simplified system can be accomplished at the speed of the interview --that is, without stopping the tape or missing part of the interview. In our experience, however, interviews are most efficiently coded from tape recordings. Efficiency is gained by having interviews available when a coder is free rather than having to coordinate the time of interviewer the consent of the respondent. However, since laws on taping telephone conversations differ from state to state, one should ascertain the state requirements for making recordings. In our studies very few respondents declined to be recorded.

The coding form in Figure 2 allows coders to enter information from a number of interviews on the same form, making tabulation of rates very easy. While we have not attempted it, coding could be accomplished with a computer and a direct data entry system. Some programming would be necessary, but rates of problem-indicating behaviors would be immediately available.

However the tabulation is done, the goal is to have coding results available to identify problems with the questions at the debriefing. The moderator uses the data as a basis for gaining interviewers' input from their experiences that will help to diagnose the nature of the problem.

Selection and Training of Coders. Two or three experienced interviewers can be selected and are easily trained to code behavior. Interviewers know the interviewing techniques specified by the organization and are familiar with pretesting. In our experience an hour or so of introduction to the questionnaire and discussion of the code categories is sufficient basis for practice coding.

Practice coding consisted of having a few interviews coded by each of the coders, with the results compared and differences discussed. In a short time coders achieve a satisfactorily high level of agreement, and production coding can begin. It is desirable to have the coders independently code a small sample of interviews as production interviewing proceeds to maintain standard interpretations over time.

Interviewer Ratings of Questions

Pretesting is usually carried out by interviewers with little or no special training for the task. In one part of this study (Chapter 4) the usual procedures were modified by providing special training to sensitize interviewers to question problems, and by having them rate each question at the completion of the interviewing for question wording problems and for several types of respondent problems. The special sensitivity training alone did not appear to lead to significant improvement in enabling interviewers to report respondent problems at the debriefing. Interviewer ratings of each question at the completion of their interviewing, however, provided a practical, systematic technique for summarizing interviewers' evaluations of questions, and for that reason they are a useful addition to pretesting procedures.

While the techniques of behavior coding and special probes are useful in evaluating questions, the interviewer role is of central importance. Our recommendations are for techniques that focus the interviewer's attention more directly on identifying question problems and the reasons for the problems. Having interviewers rate questions for problems helps to achieve this goal. As a basis for making the ratings, interviewers are instructed to make notes after each interview on a master copy of the questionnaire of any problems after each interview on a master copy of the questionnaire of any problems they or the respondent had. At the completion of their interviewing, each interviewer is asked to complete a standardized rating form. The ratings are to evaluate the types of problems they had with the questions. Space is provided for comments on possible reasons for the problems. Specifically, the interviewers rate each question on each of three types of problems:

1. Question wording problems. Problems for the interviewers in reading the questions or potential problems for respondents in understanding the language and phrasing of the question as written.

2. Understanding problems. Problems for respondents in understanding the terms used, or the ideas or concepts in the questions.

3. Response problems. Problems for respondents in answering the question due to such factors as the inaccessability of the information requested, difficulty in recalling or organizing information, and difficulty in categorizing responses into categories provided.

A three-point scale is used for rating each type of potential problem for a question:

A. No evidence of a problem.
B. Possible problem.
C. Definite problem.

A sample rating form is shown (Figure 3). In addition to making these ratings, in columns 4 and 5 on the form interviewers are encouraged to note other problems and comment on the nature or cause of problems. The notes in columns 4 and 5 are especially useful to focus the debriefing discussion on diagnosing problem sources. Aside from their use in the debriefing, the investigator can also consult the notes to supplement information from the behavior coding and special probes.

Prior to interviewing, interviewers are introduced to the types of potential problems and the rating procedures. With this background the interviewers listen to tape recordings of interviews. The tapes can be selected to illustrate a variety of problems. The discussion focuses on evidences of problems, and how the questions might be rated.

Interviewer Debriefing

It is the usual practice to hold a debriefing session following the pretest in which interviewers communicate their experiences and discuss problems with the questionnaire with an interviewing supervisor and the research staff. Chapter 2 describes some of the characteristics of debriefing sessions that present barriers to adequate evaluations of questions. The techniques recommended in this chapter are intended to provide major inputs to the debriefing. They are designed to identify major problems prior to the debriefing. The information they provide serves to introduce greater of the reasons for major question problems. By this means, interviewers can help provide diagnostic information essential for devising question revisions.

Our recommendation is that the debriefing moderator should review the behavior coding results and interviewer ratings for each question prior to the debriefing session. Interviewers should bring to the debriefing the master copies of the questionnaire on which they have recorded their comments.

The moderator, usually an interviewing supervisor, begins the session by asking for general, overall comments on the questionnaire, and then proceeds with a review of each question, focusing on those for which the behavior codes or interviewer ratings signal some difficulty. Remarks by the interviewers are focused on the causes of the problems and on suggestions for rewording questions. The moderator's job is to control the session, focusing the discussion on question problems and their diagnoses. The moderator needs to keep the discussion channeled into productive interaction by encouraging contributions from each interviewer and avoiding having one member dominate the session. Some training of the moderator in leading group discussions will assist in productive debriefings.

The result of this session and the analysis of the objective information is then available to the investigator to redesign problem questions. In some studies where substantial changes are needed, a second or even a third pretest may be necessary.

Special Probes

Special probe questions have the potential to reveal respondent problems that would otherwise go undetected because they do not lead to overt behaviors that can be coded or that interviewers can observe. For example, respondents may think they understand a question and give an adequate answer readily, but their sense of what the question means is different from what the researcher intended. Even when problems are evident through behavior coding or interviewer observation, answers to special probe questions may help to diagnose sources of difficulty.

Special probe questions may be asked to encourage respondents to elaborate on their answers, to explain how they interpreted a question, or to describe difficulties they had with a question. We found that probes aimed at question comprehension that asked respondents to define a concept or asked what they included in a response often provided valuable information about whether or not the concept was clear to the respondent and whether or not it was the one intended by the question. Neither the "tell me more" type of probe nor probes that directly asked respondents to report problems. (e.g., "Was it clear what we meant by ...?") produced useful information. Probes focusing on information retrieval and response formation were not effective in revealing problems.

We recommend using special probes to supplement the diagnostic information provided by behavior coding, interviewer ratings, and the debriefing. These probes should ascertain whether key terms and concepts are consistently understood. This type of problem often is hidden -- respondents may readily respond to questions with no indication from their behavior that a problem of understanding exists. Learning about such problems requires the investigator to predict that a problem might exist. That is, the researcher must suspect or at least wonder whether the respondent will understand or interpret correctly what is being requested.

The investigator's experience in the preliminary question development phases or past experience may raise the question of whether respondents have a common understanding. Technical terms or more abstract concepts are prime candidates for special probes. For example, possible problems may involve meanings of technical terms like HMO, clinic, doctor, chronic illness, etc., or interpretations of such concepts as trouble or worry. Effective probes need to be directed at possible sources of misunderstanding. In our experience the more specific the probes, the better the information. Thus, to learn what the respondent included as "doctors," a series of probes might ask specifically whether various medical professionals were included in his or her definition of "doctors." Special probes are incorporated most easily at the end of the questionnaire, to avoid possible influence on responses. Only a limited number of questions can be probed without unduly prolonging the interview. We recommend that about five concepts or terms that are of central importance to the study objectives and that the researcher suspects may be misunderstood be selected for special probes.

COST CONSIDERATIONS

The procedures we recommend were designed to add minimally to the time and cost of the usual pretest. The only significant cost increase is for behavior coding and for special probes. For behavior coding our recommendation is to use two or three coders. This will permit an assessment of level of coding reliability and minimize training and administrative costs. Training takes a few hours per coder. Coding time is approximately the same as the interviewing time.

While we recommended taking approximately 50 pretest interviews, that number may not be necessary to obtain adequate estimates of problems for questions with which the investigator has previous experience. Special probes added to the interview may add five to ten minutes of interviewing time. Neither of these components adds major costs to regular pretest procedures.

CONCLUSION

Question evaluation has been an unsystematic process, and not a very effective one. All the questions studied in this research had been used in major surveys, yet they contained a substantial number of serious problems. More developmental work prior to pretesting will certainly improve question design. However, laboratory studies and focus group discussions are not a substitute for testing questions in realistic data collection settings with representative respondents and interviewers. We think that improved pretest techniques, using the kind of procedures outlined above, can make a major contribution to the quality of measurement in survey research.

Research Needed

In investigating a new area of research it is not surprising that all of the questions posed in the original plan are not fully answered. Such is the case with this study. We note here some topics needing further investigation.

The behavior coding was the most successful of the techniques in supplying quantitative measure of question problems. This technique has a longer history than the others and was more easily adapted for pretest purposes. However, more experience is needed with a greater variety of questions to be more secure in generalizing to the codes and procedures to use. Our questionnaire included only a few attitude questions and a few open questions, and most of the response categories were fairly simple. Testing with a broader variety of question types and topics as well as response modes is needed for greater security in recommending procedures.

The special probes were not as productive as we had hoped. We expected that general probes that encourage respondent comments would provide a basis for understanding the nature of the question problems identified. Such probes are easy to design and use, but were generally nonproductive. Probes that were productive were harder to design and could be used with only a few questions without adding substantially to the interview length. Further investigation is needed to find some other approach to probing that will be more productive. Techniques are needed for identifying topics warranting special probes.

Interviewer identification of problems with questions is another topic needing further investigation. If interviewers could be trained to identify problem questions reliably, behavior coding would be unnecessary. Our experience (and the experience of many other investigators) is that this is difficult to achieve. Our attempts to sensitize interviewers to problems was unsuccessful. We do think, however, that use of question rating procedures that force interviewers' attention on question problems has potential to considerably improve the quality of data collection. More development and testing need to focus on what ratings interviewers can make reliably, how to train them, and how they can be taught to reliably identify the bases or reasons for respondent problems.

REFERENCES

Cantril, H.
Gauging Public Opinion. Princeton, NJ: Princeton University Press, 1944.

Payne, S.L.
The Art of Asking Questions. Princeton, NJ: Princeton University Press, 1951.

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

Sudman, S., and N.M. Bradburn
Asking Questions. A Practical Guide to Questionnaire Deisgn. San Francisco, CA: Jossey-Bass, 1982.

Figure 1. Behavior Code Categories

Interviewer Question-Reading Codes	Definition of Codes
No change	Question read as printed.
Slight change	Slight changes not affecting the meaning. Includes question readings that are not complete because the R interrupted with an answer, if the portion read contained a slight but no major change.
Major change	Change in reading the question that alters the meaning of the question or response task. Includes question readings that are not completed because the R interrupted with an answer, if the portion read included major change.
Respondent Behavior Codes	
Interrupts with answer	The R interrupts the initial question reading with answer.
Requests clarification	The R asks for a repeat or a clarification of the question, or makes a statement that indicates uncertainty about question meaning.
Qualified answer	The R gives an answer that meets question objectives but is qualified to indicate uncertainty.
Inadequate answer	The R gives an answer that doesn't meet the question objective.
Don't know	The R gives a "don't know" or equivalent answer.

Figure 2. Codesheet for Behavior Coding

	Q1	Q2	Q3	Q4	Q5	Q6	etc.
<u>Question asking</u>: No change							
Slight change							
Major change							
<u>Responses</u>: Interrupts with answer							
Requests clarification							
Qualified answer							
Inadequate answer							
Don't know							
TALLY QUESTION ASKING NO CHANGE SLIGHT CHANGE MAJOR CHANGE RESPONSES INTERRUPTS W/ANSWER CLARIFICATION QUALIFIED ANSWER INADEQUATE ANSWER DON'T KNOW							

Figure 3. Interviewer Rating Form

Columns 1-3 - Use the following code for each potential problem:

A. No evidence of Problem
B. Possible Problem
C. Definite Problem

Column 1 - Potential interviewer problems in reading question as worded or respondent problem in understanding the question as worded.
Column 2 - Potential respondent problems understanding the terms or ideas in the question.
Column 3 - Potential response problems; accessability or recall difficulties, or difficulty responding in terms of categories provided.

Column 4 - Describe other respondent or interviewer problems.
Column 5 - Comments to illuminate possible basis for problem identified in Columns 1-4.

Question Number	1 Wording Problems	2 Under-standing Problems	3 Response Problems	4 Other Problems	5 Comments
1					
2					
3					
4					
5					
6					
7					
8					
9					
etc.					

APPENDIX A

Questions and Special Probes

The left side of each page includes the original versions of the questions used in the First Pretest (Groups 1-3, Figure 1). The right side includes the revised versions used in the Second Pretest (Groups 4-5). Questions are identified by the numbers used with the original versions. Where revision of a question involved substitution of several questions, all the new questions are labeled with the same question number.

Special probes used during the interviews for Groups 2 and 4 are reproduced following the relevant questions. The lists conclude with the special probes used at the end of the interview.

Questions and Special Probes Used during the Interview

First Pretest (January, 1988)

Q2. 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 the other times you actually saw or talked to a medical doctor of any kind about your health.

Second Pretest (July, 1988)

Q2. Have you been a patient in a hospital overnight in the past 12 months since July 1, 1987?

(Not counting when you were in a hospital overnight,) During the past 12 months since July 1, 1987, how many times did you actually <u>see</u> any medical doctor about your own health?

During the past 12 months since July 1, 1987, were there any times when you didn't actually <u>see</u> the doctor, but saw a nurse or other medical assistant working for the doctor? ... (IF YES) How many times?

During the past 12 months since July 1, 1987, did you get any medical advice, prescriptions or results of tests over the telephone from a medical doctor, nurse, or medical assistant working for a doctor?... (IF YES) How many times?

3. 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?	3. 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?
4. What was the purpose of that visit?	4. SAME QUESTION
5. Where did you see the doctor this last time--at a doctor's office, a clinic, a hospital emergency room or outpatient clinic or some other place? (RECORD IF DOCTOR'S OFFICE/ PRIVATE CLINIC, HOSPITAL OUTPATIENT CLINIC, HOSPITAL EMERGENCY ROOM, OR COMPANY/SCHOOL CLINIC)	5. People see medical doctors at various places, such as at hospital emergency rooms or out-patient clinics; at company or school clinics; and at doctor's offices or private clinics. What kind of place did you go to for this last visit to a medical doctor?
6. Was that place a health maintenance organization or health care plan (that is, a place you go for all or most medical care, which is paid for by a fixed monthly or annual amount?	6. Do you belong to an HMO or health care plan that has a list of people or places you go to, in order for the plan to cover your health care costs? (IF YES) Was your last visit to a medical doctor covered by your health plan?
7. 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 or will pay for. If you don't know the exact amount, please give me your best estimate.	7. 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?
PROBE: How hard was it for you to figure out how much it is?	PROBE: How did you figure out how how much you paid?

8a. During your last visit for medical care, were you completely satisfied, somewhat satisfied, or not at all satisfied with the amount of time you had to wait to see the doctor once there?	8a. During your last visit for medical care, how did you feel about the amount of time you had to wait to see the doctor once there--were you completely satisfied, somewhat satisfied, or not at all satisfied?
8b. The quality of the care you felt was provided at that visit?	8b. How did you feel about the quality of care provided at that visit--were you completely satisfied, somewhat satisfied, or not at all satisfied?
9. When was the last time you had a general physical examination or check-up? (RECORD MONTH AND YEAR) PROBE: What was the main reason you went for that visit?	9. The next question is about a general physical examination--I mean not just to see about some problem or complaint, but a general examination. In what month and year did you last have a general physical examination or checkup? PROBE: What kinds of examinations or tests did you have done at that visit?
10. About how long has it been since you last had your blood pressure taken by a doctor or other health professional? (RECORD NBR. OF DAYS, WEEKS, MONTHS OR YEARS)	10. In what month and year did you last have your blood pressure taken by a doctor or other health professional?
10a. Blood pressure is usually given as one number over another. Were you told what your blood pressure was, in numbers?	10a. SAME QUESTION
10b. What was your blood pressure, in numbers?	10b. As close as you can remember, what was your blood pressure, in numbers?

11. During the past 12 months, since January 1, 1987, how many times did you go to a dentist, dental surgeon, oral surgeon, orthodontist, dental assistant, or any other person for dental care?

12. About how long has it been since you were last treated or examined? (RECORD IF WITHIN LAST 2 WEEKS, MORE THAN 2 WEEKS TO 6 MONTHS AGO, MORE THAN 6 MONTHS TO 1 YEAR AGO, MORE THAN 1 YEAR TO 2 YEARS AGO, MORE THAN 2 YEARS TO 4 YEARS AGO, MORE THAN 4 YEARS AGO, NEVER)

PROBE: How did you figure out when that was?

13. What did you have done during this visit? (CHECK ALL THAT APPLY IN 12-ITEM CHECK LIST)

14. The next questions are about your health <u>overall</u>. Would you say that your health in general is excellent, very good, good, fair, or poor?

PROBE: How much trouble did you have in deciding what we meant by your health in general?

11. The next question is about dental care from a dental professional such as a dentist, a dental surgeon, an oral surgeon, an orthodontist, or a dental assistant. In the past 12 months since July 1, 1987, how many times have you been to a dental professional?

12.(IF NO VISIT WITH 12 MONTHS) About how many years ago was the last time you were treated or examined for dental care?

(IF VISIT WAS WITHIN 12 MONTHS) 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?

PROBE: How did you figure out when that was?

13. SAME QUESTION

PROBE: (IF <u>X-RAYS</u> NOT CHECKED) Did you have any x-rays at that visit?

PROBE: (IF <u>CLEANING TEETH</u> NOT CHECKED) Did you have your teeth cleaned at that visit?

14. SAME QUESTION

PROBE: Could you tell me what you had in mind when you answered that?

15. Compared to other people your age, would you say your health is better than others, about the same, or worse than others? PROBE: Could you tell me more about that?	15. Compared to other people your age, would you say your health is probably better than others, about the same, or probably worse than others?
16. Are you limited in any way in any activities because of an impairment or health problem?	16. Are you limited in any activities because of an impairment or health problem? (IF NOT LIMITED) Even though you may not be seriously limited, do you have any trouble at all with any activities because of a physical problem?
16a. In what way are you limited? (RECORD LIMITATION, NOT CONDITION)	16a. What activities do you have trouble with?
17. 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? (Include days while an overnight patient in a hospital.)	17. The next question is about extra time you have spent in bed because of illness or injury (including time spent in the hospital). During the past 12 months since July 1, 1987, on about how many days did you spend several extra hours in bed because you were sick, injured, or just not feeling well?
18. In the past year, would you say you have experienced pain very often, fairly often, occasionally, or not at all? PROBE: How much trouble did you have deciding what you should include as pain?	18. SAME QUESTION PROBE: Could you tell me more about that?
19a. How much of the time, during the past month, have you been a very nervous person? Would you say 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?	19a. SAME QUESTION

19b. During the past month, how much of the time have you been a happy person? Would you say 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?

19b. SAME QUESTION

19c. How often, during the past month, have you felt so down in the dumps that nothing could cheer you up? Would you say 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?

19c. SAME QUESTION

20. Sometimes people have things they want to do but they just feel too weak, too tired, or they don't have enough energy to do them. How often do you feel this way--a lot of the time, some of the time, once in a while, or do you never feel this way?

PROBE: Could you tell me more about that?

20. Sometimes people feel too weak, too tired, or just don't have enough energy to do the things they want to do. Do you feel this way a lot of the time, some of the time, once in a while, or do you never feel this way?

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

21. In the past 4 weeks, beginning Monday (DATE 4 WEEKS AGO) and ending this past Sunday (DATE LAST SUNDAY), have you played any sports, done any hobbies involving physical activity or done any exercise, including walking?

21a. In the past 4 weeks, on how many days have you done any exercise, sports, or physically active hobbies?

21a. During the past 4 weeks, would you say you have done any of those activities 1 to 4 days, 5 to 9 days, 10 to 19 days, or 20 or more days?

22. Do you exercise or play sports regularly?

22. Do you do any sports, hobbies involving physical activity, or any exercise, including walking, on a regular basis?

PROBE: Do you not do these things at all, or do you do them, but not on a regular basis?

22a. For how long have you exercised or played sports regularly? (RECORD NBR. OF DAYS, WEEKS, MONTHS, OR YEARS)

22a. Did you begin doing these kinds of activities on a regular basis less than 6 months ago, 6 months to a year ago, between 1 and 4 years ago, or over 4 years ago?

PROBE: About how often do you do those activities?

23. Would you say that you are physically more active, less active, or about as active as other persons your age?

23. Thinking about physical activity, would you say you probably are more active, less active, or about as active as other persons your age?

23a. Is that (a lot more or a little more/a lot less or a little less) active?

PROBE: How hard was it for you to decide which one of these answers to pick?

23a. (IF MORE ACTIVE)
Is that a lot more active or a little more active than others your age?

23b. (IF LESS ACTIVE)
Is that a lot less active or a little less active than others your age?

PROBE: Could you tell me more about that?

24. How much hard physical work is required in your main daily activity? Would you say a great deal, a moderate amount, a little, or none?

PROBE: Could you tell me more about that?

24. SAME QUESTION

24a.(IF GREAT DEAL OR MODERATE AMOUNT) About how many hours per day do you perform hard physical work in your main daily activity?	24a. (IF GREAT DEAL OR MODERATE AMOUNT) Would you say you do physical work in your main daily activity for less than 2 hours a day, 2 to 4 hours, or over 4 hours?
25. How many days a week do you think a person should exercise to strengthen the heart and lungs?	25. About how many days a week do you think a person needs to exercise, to strengthen the heart and lungs? PROBE: Would you tell me more about your thinking on that?
26. For how many minutes do you think a person should exercise on each occasion so that the heart and lungs are strengthened?	26. About how many minutes do you think a person needs to exercise each time, to strengthen the heart and lungs?
27. (During those (NUMBER IN Q26) minutes) How fast do you think a person's heart rate and breathing should be to strengthen the heart and lungs? Do you think that the heart and breathing rate should be --no faster than usual, a little faster than usual, a lot faster but talking is possible, so fast that talking is not possible?	27. When a person exercises to strengthen the heart and lungs, would you say the heart rate and breathing should be no faster than usual, a little faster than usual, or a lot faster than usual? (IF A LOT FASTER THAN USUAL) Would that be a lot faster but talking is possible, or so fast that talking is not possible? PROBE: Could you tell me more about that?
28. About how much do you weigh without shoes?	28. SAME QUESTION
29. Do you consider yourself overweight, underweight, or just about right?	29. SAME QUESTION

30a. We are interested in how well people take care of themselves. Do you think you do very well, fairly well, or not so well as far as eating nutritious meals?	30a. We are interested in how well people take care of themselves. How well do you think you do as far as eating nutritious meals--very well, fairly well, or not so well?
30b. (Do you think you do very well, fairly well, or not so well as far as) keeping at the right weight?	30b. How about keeping at the right weight--do you think you do very well, fairly well, or not so well?
PROBE: How hard was it for you to pick an answer that describes how you really felt?	
30c. Taking care of your teeth or dentures?	30c. How well do you think you do as far as taking care of your teeth or dentures--very well, fairly well, or not so well?
31a. In answering the following questions, think about your eating pattern over the last 12 months. For each food group I mention, please tell me the average number of days each week you have eaten this type of food. How often do you have red meat, such as beef, pork, lamb, liver, and so on?	31a. In answering the following questions, think about your eating pattern over the last 12 months. For example, what is the average number of days each week you have red meat, such as beef, pork, lamb, liver, and so on?
32a. What is the number of servings on a typical day?	32a. On days when you eat red meat, how many servings do you usually have?
	PROBE: Would you include things like bacon, hot dogs, or lunch meats as red meat?
31b. What is the average number of days each week you have eggs?	31b. SAME QUESTION
PROBE: How hard was it for you to figure that out?	
32b. What is the number of servings on a typical day?	32b. On days when you eat eggs, how many eggs do you usually have?

31c. What is the average number of days each week you have butter?

32c. What is the number of servings on a typical day?

PROBE: Could you tell me more about that?

33. I am going to read two statements. Please tell me which one you agree 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.

PROBE: Could you tell me more about that?

34a. 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, cigarette smoking? (Do you think it definitely increases, probably increases, probably does not increase, or definitely does not increase a person's chances of getting heart disease?)

31c. The next question is just about butter. Not including margarine, what is the average number of days each week you have butter?

PROBE: Were you counting butter that was used in cooking and baking, or not?

32c. On days when you eat butter, how many servings do you usually have?

PROBE: Could you tell me more about that?

33. How much effect do you think what people eat and drink has on whether they develop major diseases--would you say it has a large effect, some effect, little effect, or no effect?

PROBE: Could you tell me more about that?

34a. The next questions are about things that may or may not affect a person's chances of getting heart disease. How much effect do you think cigarette smoking has on whether a person will get heart disease--do you think it has a large effect, some effect, little effect, or no effect?

34b. high blood pressure?

34b. How much effect do you think high blood pressure has on whether a person will get heart disease--do you think it has a large effect, some effect, little effect, or no effect?

34c. diabetes? (Do you think it definitely increases, probably increases, probably does not not increase, or definitely does not increase a person's chances of getting heart disease?)

34c. What about the effect of diabetes on getting heart disease--(do you think it has a large effect some effect, little effect, or no effect?)

34d. being *very* overweight?

34d. How much effect do you think being *very* overweight has on whether a person will get heart disease--(do you think it has a large effect, some effect, little effect, or no effect?)

34e. drinking coffee with caffeine?

34e. How much effect do you think drinking coffee with caffeine has on whether a person will get heart disease--(do you think it has a large effect, some effect, little effect, or no effect?)

34f. eating a diet high in animal fat?

34f. How much effect do you think eating a diet high in animal fat has on whether a person will get heart disease--(do you think it has a large effect, some effect, little effect, or no effect?)

34g. high cholesterol?

PROBE: How much trouble did you have deciding what we meant by getting heart disease?

34g. What about the effect of *high cholesterol* on getting heart disease--(do you think it has a large effect, some effect, little effect, or no effect?)

35. What do you think are the warning signs or symptoms of cancer? (CHECK ALL THAT APPLY IN 14-ITEM CHECKLIST)	35. What are some of the symptoms a person should be concerned about because they may be warning signs of some kind of cancer? (CHECK ALL THAT APPLY IN 14-ITEM CHECKLIST)
36. Where do you get most of your most of your information about AIDS? (CHECK ALL THAT APPLY IN 19-ITEM CHECKLIST) PROBE: What kind of difficulties did you have in answering this question?	36. SAME QUESTION
37a. Here are methods some people use to prevent getting the AIDS virus through sexual activity. After I read each one, tell me whether you think it is very effective, somewhat effective, not at all effective, or if you don't know how effective it is in preventing getting the AIDS virus through sexual activity. How effective is using a condom?	37a. One way to get the AIDS virus is from sexual activity. We want to ask how effective you think some methods are for preventing getting the AIDS virus <u>from sexual activity</u>. How about using a condom--would you say it is very effective, somewhat effective, not at all effective, or don't you know how effective it is in preventing getting the AIDS virus from sexual activity?
37b. Being celibate, that is, not having sex at all?	37b. Would you say being celibate, that is, not having sex at all, is very effective, somewhat effective, not at all effective, or don't you know how effective it is in preventing getting the AIDS virus from sexual activity?
37c. Two people who do not have the AIDS virus having a completely monogamous relationship, that is, having sex <u>only</u> with each other?	37c. How about two people who do not have the AIDS virus having a completely monogamous relationship--that is, having sex <u>only</u> with each other? Would you say that is very effective, somewhat effective, not at all effective, or don't you know how effective it is in preventing getting the AIDS virus from sexual activity?

Special Probes Used at the End of the Interview

NOTE: Probes are identified by the number of the relevant question. Some questions had more than one probe. Multiple probes for a question that appeared in the same questionnaire form are grouped under one question number. Probes of the same question that appeared in another form are separately grouped and labeled with the question number.

(INTRODUCTION FOR GROUP 2 INTERVIEWS)
That's all the regular questions. Now, I'd like to ask you about some of the questions we just asked you. We think some of them might be hard to understand or hard to answer. It's important in our research to be sure we are asking questions that are easily understood and can be answered without too much trouble.

2. The last question we want to ask you about is how many times in the past 12 months you saw or talked with a doctor or assistant about your health. In answering that question, if you had gotten medical advice on the telephone from a doctor or assistant, would you have counted it in your answer?

We're interested in what people include as doctors or assistants. When you think of a doctor or assistant---
Would you include a chiropractor or not?
Would you include a physical therapist or not?
Would you include a podiatrist or not?
Would you include an optometrist or not?
Would you include a psychiatrist or not?
Would you include a nurse or not?
Would you include a laboratory or x-ray technician, or not?

(INTRODUCTION FOR GROUP 4 INTERVIEWS)
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 earlier and get some of your thoughts about them.

2. One question said, "During the past 12 months since July 1, 1987, how many times did you actually see any medical doctor about your own health?" We're interested in what people include as medical doctors. When you think of a medical doctor--- Would you include an osteopath or not?
Would you include a dentist or not?
Would you include a psychiatrist or not?
Would you include a dermatologist or not?
Would you include an opthalmologist or not?

We also asked you about times in the last year that you got medical advice, prescriptions, or test results over the telephone from a medical doctor, nurse, or other medical assistant working for a doctor. You said this happened (NUMBER) times. We think this may be a difficult question to answer. Would you say the information you gave was pretty accurate, a rough guess, or what?

2. One question said, "During the past 12 months since July 1, 1987, how many times did you actually see any medical doctor about your own health?" We're interested in what people include as medical doctors. When you think of a medical doctor ---

Would you include a chiropractor or not?

Would you include a physical therapist or not?

Would you include a podiatrist or not?

Would you include an optometrist or not?

Did you see any of those kinds of people during the last year?... (IF YES) Did you include them in the visits you told me about earlier?

Did you see any of these kinds of of people during the last year?... (IF YES) Did you include them in the visits to medical doctors you told me about earlier?

2. For example, we asked you this question, "During the past 12 months, since, Jan. 1, 1987, how many times have you seen or talked with a doctor or assistant about your health?" In answering that question, how hard was it for you to figure out the number of times you saw or talked to a doctor--was it very hard, somewhat hard, or not hard at all?...Tell me more about that.

2. One question said, "During the past 12 momths since July 1, 1987, how many times did you actually see any medical doctor about your own health?" We think it might be difficult for people to remember the number of times. Could you tell me about any problems you might have had figuring out the answer?...Do you have any (other) comments about that question?

3. For example, one question was, "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?" In answering that question, how did you figure out when the last time was?

7. For example, we asked about your most recent doctor's visit. The question was "How much did you pay, or will you have to pay, out of pocket, for your most recent visit?"

7. We asked you about how much it cost you or your family for your most recent visit to a medical doctor. The question said, "Not including what insurance pays,

How hard was it for you to figure out those out-of-pocket costs--would you say it was very hard, somewhat hard, or not hard at all?...Tell me more about it.

about how much did you pay or will you pay for the visit?" We think it may be difficult for people to figure out or remember the cost. Did you have any difficulty at all? ... (Could you tell me more about that?)

8b. We also asked this question: "During your last visit for medical care, were you completely satisfied, somewhat satisfied, or not satisfied at all with the quality of the care you felt was provided at that visit?" How hard was it for you to pick an answer that describes how you really felt?"

10.Another question we asked was "<u>About</u> how long has it been since you <u>last</u> had your blood pressure taken by a doctor or other health professional? Do you think you answer was exact, pretty close, or not very close to the actual time?

10. Another question we asked was, "In what month and year did you <u>last</u> have your blood pressure taken by a doctor or other health professional? Do you think your answer was exact, pretty close, or not very close to the actual time?

10b. (For example, one/Another) question was "What was your blood pressure, in numbers?" You said your blood pressure was (ANSWER). Do you think your answer was exact, pretty close, or not very close?

16. Finally, we asked whether you are limited in <u>any way</u> in any activities because of an impairment or health problem. You said you were (limited/ not limited). Are you limited in any (other) way <u>at all</u> in what you can do, because of health reasons?

Are there any things in your work, recreation, or social life that you <u>can't do as well as you would like</u>, because of health reasons?

Are there any things in your work, recreation, or social life that you've

16. (IF LIMITED) Another question asked whether you are limited in any activities because of an impairment or health problem. You said you were limited. Besides what you told me, are you limited in any other way <u>at all</u> in what you can do, because of health reasons?

(IF NOT LIMITED) Another question asked whether you are limited in any activities because of an impairment or health problem. You said you were <u>not</u> limited. Are you limited in any way <u>at all</u> in had to give up doing, because of health reasons?

17. (For example, one/Another) question asked about stays in bed. We asked, "During the past 12 months, that is, since Jan. 1, 1987, about how many days did illness or injury keep you in bed more than half of the day?" How clear was it to you what to include as illness or injury?

How clear was it to you what to include as a half day in bed?

17. Finally, we asked the question, "During the past 12 months, that is, since Jan. 1, 1987, about how many days did illness or injury keep you in bed more than half of the day?" Besides what you told me, were there any other times during the year that you stayed in bed late, went to bed early, or just lay down because you weren't feeling well?...(IF YES) Would you say that (this time/any of these times) lasted a half-day or longer?

Were there any other times during the year when you were not in bed but were lying down for half a day or longer because you just weren't feeling well?

We're interested in what people include as illness. Some people lie in bed for half a day or more because they just feel tired. Would you count that as staying in bed because of illness?

what you can do, because of health reasons?

Are there any things in your work, recreation, or social life that you can't do as well as you would like, because of health reasons?

Are there any things in your work, recreation or social life that you've had to give up doing, because of health reasons?

17. Another question was about times in the last year that you spent several extra hours in bed because you were sick, injured, or just not feeliing well. We're not sure whether it's clear what we meant by being "sick, injured, or just not feeling well." Did you have any trouble deciding what we meant by that?...(Could you tell me more about that?)

17. We also asked you about extra time you spent in bed because of illness or injury. The question said, "During the past 12 months since July 1, 1987, on about how many days did you spend several extra hours in bed because you were sick, injured, or just not feeling well?" Were there any times during the year when you were not actually in bed but were lying down for several hours because you were sick, injured, or just not feeling well?...(IF YES) Did you include those times in the answer you gave me earlier?

We're also interested in what people include as being sick or injured or just not feeling well. Some people spend extra time in bed because they just feel tired. Would you count that as staying in bed because of being sick, injured, or just not

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?	feeling well? 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 being sick, injured, or just not feeling well?
19b. We asked the question, "During the last month, how much of the time have you been a happy person? Would you say 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?" In answering that question, how hard was it for you to pick an answer that describes how you really felt?	
21. (IF EXERCISE) We also asked about physical exercise. 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?	21. (IF EXERCISE) We also asked about physical exercise. You said that in the past 4 weeks you had done some exercise, sports, or hobbies involving physical activity. Could you tell me more about that?
(IF NO EXERCISE) We also asked about physical exercise. You said that in the past 4 weeks you had not done any exercise, sports, or physically active hobbies. Did you get any exercise <u>at all</u> during that time?	(IF NO EXERCISE) We also asked about physical exercise, sports or hobbies involving physical activity. Did you get any exercise <u>at all</u> during that time?
	21a. Earlier you told me that you had done some exercise, sports, or hobbies involving physical activity on (# OF DAYS REPORTED) during the past four weeks. Could you tell me how you figured out your answer?
27. We also asked about strengthening the heart and lungs through exercise. The question was, "Do you think that the heart and breathing rate should be --no faster than usual, a little faster than usual, a lot faster but talking is possible, so fast that talking is not possible?" How hard was it for you to decide which one of these	27. We also asked about strengthening the heart and lungs through exercise. One question was about how much faster you think the heart rate and breathing should be when a person exercises to strengthen the heart and lungs. The answer you gave was (ANSWER). Could you tell us a little bit

answers to pick?

34d. In another question we asked how much being <u>very</u> overweight affects a person's chances of getting <u>heart disease</u>. That is, do you think it definitely increases, probably increases, probably does not, or definitely does not increase a person's chances of getting heart disease? In answering that question, how hard was it for you to decide which one of those answers to pick?

36. We also asked a couple of questions about AIDS. One was, "Where do you get most of your information about AIDS?" What kinds of difficulties did you have in answering this question?

about how you decided on that answer?

31b. Another question asked about the average number of days each week you have eggs. We'd like to know how people figured out their answers. When you answered that question, were you including times you ate eggs used in baked goods and cooking, or not?

APPENDIX B

Debriefing Coding Procedures for Groups 1 and 2

The category of respondent problems (other than interruptions) contains a wide range of problems for which a variety of interviewer statements (such as suggestions, descriptions of problems, and examples of the behavior of single respondents) had to be considered. In order for a question to be judged as problematic, we decided that the following two types of evidence had to be given at the debriefing:

Main evidence
An interviewer had to describe a problem in a way that suggested that more than one of her respondents behaved in a way used as a problem indicator by the behavior analysis system (i.e. by using a plural pronoun in the description of the number of respondents with the problem). One interviewer's main evidence could be cancelled out by another who stated that none of her respondents had the problem.

Supporting evidence
A suggestion for problem resolution, an agreement that the problem existed or that the suggestions were useful, a further statement that could be classified as "main evidence", or at least two examples of the behavior of individual respondents was considered as supporting evidence that problems with a question occurred frequently or were considered important by interviewers.

As the scheme shows, interviewers' subjective assessments or hypotheses about a question, taken alone, were not a sufficient basis for judging it to be problematic. Interviewers had to report that respondents actually behaved in ways indicating problems with the question. This requirement meant that the debriefing results could be compared easily to behavior coding results to investigate how interviewers reported actual interview experience. Two respondents may appear to be a small number on which to base a judgment, but it reflects the low frequency of interviewer reports about respondent behavior at the debriefing.

Directions for Coding Behavior with the "Pretest" Coding Form

The purpose of the Pretest study is to provide evidence about the quality of question wording as demonstrated by interviewer behavior in asking questions and by respondent behavior in answering them. Only behaviors that indicate potential problems with questions are be coded. The only interviewer Those behaviors that do not relate to problems with behavior to be coded is the initial question reading. Nearly all respondent behavior is coded.

Behavior Codes

Definitions of the behavior codes follow. Details about their application are given in the Appendix.

Question-reading codes

Question reading codes

E-Exact

Interviewer reads the question exactly as printed. It is acceptable to use a contraction or to replace a contraction with the appropriate words.

S-Slight Change

Interviewer reads the question changing a minor word (or words) that do not alter question *meaning*.

M-Major Change

The interviewer changes or rewords the question such that the meaning of the question is altered.

B-Break off

The interviewer stops and does not resume reading the question, because the respondent has interrupted.

Respondent codes

1-Interruption of initial question reading with an *answer*

Code all *answers* that interrupt the initial question reading here, regardless of whether or not the interviewer resumes reading the question.

3-Request for def/clar or RQ

R asks for repeat of all or part of the question, or R asks for clarification of the meaning of the question or for definition of a phrase or word in the question.

5-Adequate answer

Gives an answer that meets the question objectives.

6-Qualified adequate answer

Gives an answer that meets the question objectives, but that is qualified to indicate uncertainty.

7-Inadequate answer

Gives an answer that does not meet the question objectives.

8-Don't know

Gives a "don't know," or equivalent response.

9-Refusal to answer Q

Overview of the Coding System

In asking the question and arriving at an answer, the interviewer and the respondent take turns speaking. The shortest number of speaking turns is two--for example:

The interviewer reads the question exactly--Turn 1, coded E
The respondent answers adequately--Turn 2, coded 5

If the interviewer followed with a long feedback statement or a repeat of the answer, it would be turn 3, but not coded. Sequences of two or three turns are common. Here is an example of a sequence involving more turns--

The interviewer reads the question with a slight change--Turn 1, coded S
The respondent asks about the meaning of the question--Turn 2, coded 3
The interviewer repeats the question--Turn 3, not coded
The respondent says he doesn't know the answer--Turn 4, coded 8
The interviewer probes--Turn 5, not coded
The respondent gives an inadequate response followed by an adequate response--Turn 6, coded 5,7

Coding the Interview

Coding will be done from tape recordings of the interviews, using the Pretest Coding Form. Each interview will be coded on a separate form. Before beginning coding, you should enter the following information in the designated boxes on the first page of the form:

- Interviewer ID
- Coder ID
- Prime or check coding
- Log number of the interview. The last two items also should be entered on the remaining pages of the coding form.

At the end of the interview, you should enter respondent information in the designated boxes on the first page of the coding form. This information is obtained from responses to Qs 38, 39, 40, 40a and 40b, using the code values indicated in the questionnaire.

For each question asked, you will be entering codes for certain interviewer and respondent behaviors. The coding form is arranged so that all codes for a question are entered on a single line. Each question on the questionnaire is identified in the shaded column. (The left-most column of numbers is for data processing, and you should ignore it.)

Coding Question Reading

The interviewer question-reading turn is coded by checking one of the boxes labeled:

- E - exact reading
- S - slight change
- M - major change
- B - break off

Coding Respondent Behavior

Code respondent behaviors (codes 1 through 9) in the columns labeled "Respondent Behaviors." Codes for the first respondent turn are entered in the first column; codes for the second respondent turn are entered in the second such column, and so on.

Coding respondent behavior within a turn

For a given turn, you should code behaviors as they occur, with the restriction that the same code number should not be used twice in a row. Most of the time you will have no trouble identifying behaviors to be coded separatey. When you are in doubt, the general guideline is to code the behavior.

Respondent Behaviors that are not Coded

Some types of respondent behaviors will not be coded, because they do not indicate anything about difficulties with the original question. If these behaviors are the only respondent behaviors in the turn, you should ignore the entire turn.

- Respondent digressions should not be coded.

- The interviewer has repeated the respondent's answer and the respondent confirms it ("Yes," "Uh-huh"). This behavior sequence is usually a time filler, indicating nothing about the question-answering process, and the respondent turn should not be coded. The exception is, however, when it is clear to you that the interviewer repeated the respondent's answer as a probe rather than as a time filler, or if the respondent takes the opportunity to say something more than a simple affirmation, you should code the respondent behavior.

- Do not code respondent queries about the meaning of interviewer probes or statements that are not about the original question. If, however, the interviewer statement is a clarification of the original question, respondent queries should be coded.

- Respondent comments and questions about any aspect of the survey or interview other than the immediate question should not be coded.

Identifying Turns.

Most of the time you will have no difficulty identifying the end of one turn and the beginning of the next turn. In general, the end of one turn and the beginning of the next turn is when one party stops speaking and the other begins speaking.

In identifying turns you should ignore minor, casual comments by either the interviewer or the respondent such as "Uh-huh," "I see," etc. Also ignore probes or comments that one party begins to make but that the other party interrupts or ignores, and continues speaking. Finally, we do not want to include problems of hearing. You should ignore any requests for repeats of questions, answers, probes, etc., when the problem is clearly one of hearing (rather than understanding).

Annotated Questionnaire

The annotated questionnaire includes information to help you in coding. For questions preceded by introductions, we have indicated what is to be included as part of the question in judging question reading. For some questions the adequacy or inadequacy of certain types of responses may be unclear to you. To help you in coding, we have labeled such "hard to code" responses as either adequate, qualified adequate, or inadequate, on a question-by-question basis.

Appendix

Here are some guidelines for applying the behavior codes.

Question-reading codes

E-Exact

•Parenthetical material must be read completely or not at all.
•Brief transitions (And, Next, How about, etc.) are acceptable.
•For a question in a series of closed questions with the same response choices, it is acceptable to read the response choices even when they are not printed for the question (for example, Q34b).
•Question readings that the respondent has interrupted should be coded here if the interviewer perseveres and completes and exact reading.
It is acceptable for the interviewer to interject a comment explaining that she must read the question completely.)

S-Sight change

•Use this code when the interviewer misreads the question but immediately corrects the mistake and delivers the question correctly.

M-Major

•Incomplete reading of the question is coded as a major change, unless the Breakoff code applies.
•Included here are question readings where the interviewer turns the question into a statement based on previously obtained information.

Respondent codes

3-requests def/clar or RQ

•It can be an interruption of the initial question reading.
•Also code statements indicating that the R is unsure about the meaning of the

5-adequate answer

•For closed questions this means in terms of the response choices offered.
•For open questions with precodes thismeans in terms of the precodes. (The presence of "other, specify" for some questions with precodes means that a response need not fit neatly into one of the precodes to be considered adequate.)

8-don't know

•Code all "don't know" type answers here, even for questions with a DK box. If the question includes a DK response option, like Qs 37a-c, DK should be coded 5 (adequate response).

P.468264
2/88

Today's Date: ____________________

PRETEST CODING FORM

Sample

RESPONDENT

Birth Yr □□ Sex □ Education: YRS □□ HSD □ CD □

Iwer ID: □□□

Coder ID: □

Prime=1
Check=2 □ Log Number: □□□□

		Q READING				RESPONDENT BEHAVIORS 1 = interrupts Q with answer 3 = requests def/clar 5 = adequate answer 6 = qualified adequate answer 7 = inadequate answer 8 = don't know 9 = refusal to answer						
	Q #	E	S	M	B							NOTES
01	Q2											
02	Q3											
03	Q3a											
04	Q4											
05	Q5											
06	Q5a											
07	Q5b											
08	Q5c											
09	Q6											
10	Q7											
11	Q8a											
12	Q8b											

APPENDIX C
Sample Coding Form

APPENDIX D
Sample Question Rating Form for
Groups 3 and 5

PRETEST METHODOLOGY . WAVE I (PROJ# 288)

INTERVIEWER RATING FORM

Use the following code for each potential problem:

A. No evidence of Problem
B. Possible Problem
C. Definite Problem

Sample

Column 1 should be used for potential problems due to your having trouble reading the question as written.

Column 2 should be used for potential problems due to respondents not understanding words, or ideas in the question.

Column 3 should be used for potential problems due to respondents having difficulties knowing the accurate answer.

Column 4 should be used for potential problems due to respondents having trouble answering in the terms required by the question.

Question Number	Hard to Read	R has problem Understanding	R has' No Info/ No Recall	Problem w/terms	Other Problems	Comments
Q2. #times saw Dr. in past 12 months?						
Q3. Last time saw Dr.?						
Q4. Purpose of visit?						
Q5. Where saw doctor?						
Q5a.What type of clinic?						
Q5b.Type of hospital facility						
Q5c.Type of place?						
Q6. Was place HMO?						
Q7. Amount pay out of pocket?						