NEW QUANTITATIVE TECHNIQUES FOR PRETESTING SURVEY QUESTIONS

Charles Cannell*, Floyd J. Fowler#, Graham Kalton*, Lois Oksenberg*, and Katherine Bischoping*

*Survey Research Center, The University of Michigan, Ann Arbor, Michigan, 48106, U.S.A.
#Center for Survey Research, University of Massachusetts, Boston, Massachusetts, 02116, U.S.A.

Key Words: Pretests; Pilot surveys; Debriefing; Behavior coding; Probe questions; Questionnaire Design

1. 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 generally goes through several developmental stages between the researcher's formulation of objectives or hypotheses and data collection. These stages are described in detail in DeMaio (1983). Sometimes, when the researcher is exploring an unfamiliar topic, some unstructured design work may be conducted, perhaps using depth interviews, group discussions, or free-flowing interviews on which to base a first draft of the questionnaire (see Morton-Williams, 1978, for a good description of these procedures). The initial version of the questionnaire is commonly reviewed by the researcher and a few colleagues to determine whether they think that it will obtain the data required for the survey. They try to gauge whether interviewers will be able to handle the questions with ease, whether respondents will be able to comprehend what is wanted, and whether respondents will be willing and able to provide the answers. After
necessary revisions have been made, the questionnaire is then usually subjected to a more formal pretest.

The purpose of the pretest is to determine whether the individual questions and the questionnaire as a whole work as intended. Problems with individual questions may arise for a variety of reasons including the following:

(a) Questions that are difficult to ask. Interviewers may have problems reading a question because it has a complex sentence structure, or because it contains tongue twisters or words that are difficult to pronounce.

(b) Question comprehension. Comprehension problems may occur because of the use of vocabulary that is too difficult for the respondent, because the sentence structure is too complex, or because the question does not specify clearly what information is needed or the form in which it is to be reported.

(c) Lack of a common understanding of terms and concepts. Lack of a common understanding occurs when the terms or concepts used in a question are understood differently by different respondents, or the question is not interpreted as the researcher intended.

(d) Difficulty in cognitive processing of information. This common difficulty arises when the respondent is unable or unwilling to exert the level of effort needed to provide an adequate answer. Sometimes the information is simply inaccessible to the respondent. Often, the information is accessible, but the effort needed to retrieve and process it is greater than the respondent willing to make.

Besides checking on individual questions, the pretest also assesses the questionnaire as a whole. Does the questionnaire flow smoothly from one topic to the next? Do the skip patterns work as intended? Does the juxtaposition of certain questions cause problems? Even if questions work well individually, they may not do so in combination.

The procedures for pretests are well described by Converse and Presser (1986). The most common strategy is to have interviewers administer the questionnaire to some 25 to 75 respondents using standard interviewing techniques. Experienced interviewers are often chosen for pretesting. They are instructed to watch out for problems that respondents appear to have in answering questions, but are otherwise given no special training or instructions. After the pretest interviews are completed, the interviewers are debriefed either individually or, more commonly, in a group debriefing session in which the questionnaire is reviewed question by question with interviewers reporting any problems they thought had occurred.
Although the pretest plays a critical role in questionnaire development, little research has been conducted in pretesting methods. Since many problem questions filter through into final questionnaires, it appears clear that the current unsystematic methods of pretesting are not as effective as they need to be. The purpose of the research reported in this paper is to develop improved, more systematic, methods of pretesting that can be used in regular survey pretests with little or no increase in cost.

Three techniques for improving the effectiveness of pretesting in detecting problem questions are examined in this research. They are:

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

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

3. **Special training of pretest interviewers.** Pretest interviewers are given special training in how to recognize problems with questions and to identify the nature of the problems. They are also asked to rate each question with regard to the presence of different types of problems.

None of these techniques is new, each of them having been used, or suggested, for evaluating questions or for other purposes. None of them alone will identify all types of problems with survey questions. The aim of the current research is to develop and integrate the techniques into a unified and efficient pretesting procedure.

This paper reports some results from an experiment conducted to study the use of these three techniques in pretesting and to compare these techniques with the pretest debriefing procedures commonly used at present. The next section describes the study design. Sections 3, 4 and 5 present some results obtained by from the three techniques. Section 6 reviews the findings and offers some tentative conclusions on improving pretesting procedures.

2. **Study Design**

An experimental study was conducted to examine the effectiveness of the three techniques described above. There were two phases to the experiment.

A questionnaire of 60 questions was constructed for the first phase of the experiment. The questions spanned the range of topics commonly covered in health surveys, including questions on: visits for medical and dental care; health care plan membership; various aspects of health status; foods eaten; exercise taken; and knowledge concerning cancer and heart disease.
risks, and the transmission of AIDS. The questions were taken from questionnaires used for some major health surveys. The questionnaire was administered by telephone to three groups of respondents. Samples of telephone numbers were drawn by the same probability sampling procedures from lists of telephone subscribers in southeastern Michigan for each of the three groups. For cost reasons, the selection of respondents at selected telephone numbers was by a nonrandom procedure; the procedure aimed to give a balance between male and female respondents and between respondents of different ages.

The first group of respondents was interviewed using standard pretest procedures. Interviewers used regular interviewing techniques. They did not attempt to vary the questions or use extra probes. Six interviewers each took ten interviews for a total of 60 respondents.

The second group comprised 100 respondents who were interviewed with the same basic questionnaire. However, for this group special diagnostic probes that aimed to identify problems with the question were added to the questionnaire. Some probes were unobtrusive and were included during the interview; others had a more intensive focus on special issues and were added at the end of the regular interviews. Eight interviewers were used, all different from those used for group 1.

The third group of respondents was interviewed using the basic questionnaire by interviewers specially trained and given practice in identifying respondent and interviewer problems with questions. Five interviewers (all different from those used in groups 1 and 2) conducted 50 interviews in this group.

The interviews for each group were tape-recorded (with the permission of the respondents). All the interviews in groups 1 and 3 were behavior coded from the tape recordings. The interviewers from the three groups were debriefed in separate debriefing sessions after the interviews had been completed. The debriefing sessions were tape-recorded and the tape recordings for groups 1 and 2 were subjected to a content analysis.

The various analyses conducted at the end of the first phase of the experiment identified a number of questions with significant problems. These questions were revised in an attempt to remedy, or at least reduce, the problems. The revised questionnaire was then tested in the second phase of the experiment. The second phase comprised two groups, group 4 corresponding to group 2 and group 5 corresponding to group 3 of the first phase. The second phase interviews were tape-recorded, and the tape recordings of group 4 were behavior coded.

3. Behavior Coding

The behavior coding scheme employed in this study was adapted from coding schemes used in previous studies that coded behavior in survey interviews (Cannell, et al., 1975; Oksenberg, 1981; Morton-Williams, 1979; Morton-Williams and Sykes, 1984). The current coding scheme was developed with the aim of capturing aspects of the interviewer-respondent interaction.
that indicate interviewer or respondent problems with a question, and if possible what the causes of the problems might be. The coding scheme focused primarily on respondent behavior. The only interviewer behavior coded was whether the initial reading of the question was exactly as printed, had a slight change of wording that did not alter the meaning of the question, had a major change that altered the meaning, or whether the interviewer did not complete the reading of the question because the respondent interrupted. Respondents’ behaviors were coded as: interrupts with an answer; asks for clarification or for the question to be repeated; provides an adequate answer; provides a qualified adequate answer; provides an inadequate answer; gives a "don't know" response; or refuses to answer the question. Respondents' behaviors were coded as: interrupts with an answer; asks for clarification or for the question to be repeated; provides an adequate answer; provides a qualified adequate answer; provides an inadequate answer; gives a "don't know" response; or refuses to answer the question. Codes were assigned both for respondents' initial responses to a question and for their subsequent responses (i.e., after intervening interviewer behavior such as repeating the question, providing clarification, or asking a probe question). Three coders were extensively trained in how to perform the behavior coding.

Many different indicators of problems with questions can be derived from the behavior codes. After analyzing a number of these indicators and their interrelationships, the following ten indicators were chosen to tap the range of problems that behavior coding might identify:

1. Interviewer read the question with slight changes (abbreviated to "slight changes").
2. Interviewer read the question with major changes or did not complete reading the question (major changes).
3. Respondent interrupted the question reading with an answer (interruption).
4. Respondent requested clarification, explanation, or repeat of the question (clarification).
5. Respondent gave a qualified adequate answer (qualified answer).
6. Respondent gave an inadequate answer (inadequate answer).
7. Respondent gave a "don't know" answer ("don't know").
8. Respondent refused to answer the question (refusal).
9. Respondent behavior indicated some problem, that is, one or more of categories 3, 4, 5, 6, 7, and 8 (respondent problem).
10. Respondent never gave an adequate answer (no adequate answer).

A sample check on the coding of these indicators showed acceptable levels of reliability, with kappa coefficients of over 0.7 for all but the "qualified adequate answer" indicator.
These indicators were tabulated for each question, and the tabulations were used to diagnose the existence of problems. Table 1 displays the percentages of times these indicators were assigned for a selection of questions for the sixty interviews of group 1, the group interviewed with regular interviewing techniques without the special probes. As can be seen from the table, none of the questions is problem-free. Indeed, none of the questions in the whole questionnaire scored zero on all the problem indicators. Small values on the indicators are to be expected with good questions, and may be tolerated. Large values, however, indicate that a sizeable proportion of respondents have a problem with the question, and suggest that the question needs revision. It is a matter of judgment as to what constitutes "large" or "small," and in fact no definite classification is needed. On a pragmatic basis, we adopted a value of about 15% as the threshold between "small" and "large". Even if an indicator falls below the threshold, however, a revision of a question is warranted if the problem with the question can be readily identified and easily resolved.

Table 1: Behavior code indicators for a selection of questions: Percentage of times the indicator was assigned to the question in the 60 interviews of group 1.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Question number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>1. Slight changes</td>
<td>8</td>
</tr>
<tr>
<td>2. Major changes</td>
<td>19</td>
</tr>
<tr>
<td>3. Interruption</td>
<td>35</td>
</tr>
<tr>
<td>4. Clarification</td>
<td>3</td>
</tr>
<tr>
<td>5. Qualified answer</td>
<td>3</td>
</tr>
<tr>
<td>6. Inadequate answer</td>
<td>8</td>
</tr>
<tr>
<td>7. Don't know</td>
<td>0</td>
</tr>
<tr>
<td>8. Refusal</td>
<td>0</td>
</tr>
<tr>
<td>9. Respondent problem</td>
<td>48</td>
</tr>
<tr>
<td>10. No adequate answer</td>
<td>30</td>
</tr>
</tbody>
</table>

Using these guidelines, it can be seen from Table 1 that only one question, question 4, is fully acceptable. This was an open question about the respondent's last doctor visit: "What was the purpose of that visit?" Each of the other questions is subject to sizeable percentages of problems, as is clear from the summary indicator 9 that reflects all types of respondent problems. The pattern of the indicators differs between questions, suggesting that there are different sources of problems in the different questions.

Question 3 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?" This question was interrupted by respondents giving answers before the reading was completed.
in more than a third of the interviews. There were also many major changes in the interviewers' readings of the question, most of which were because the interviewer failed to complete the reading. As can be seen from the question, its structure appears to encourage respondents to interrupt with answers after the initial question has been posed, or as soon as the response choice that applies to them has been read. This kind of question structure occurred in several other questions also, and these questions had similar behavior coding results.

Question 12 was concerned with dental care. It read "About how long has it been since you were last treated or examined?" Nearly a third of respondents requested clarification, and the same proportion gave inadequate answers to the question. Three potential sources of problems were diagnosed. First, question 12 followed a question asking for the number of visits for dental care in the past year. Although these two questions are logically distinct, the sequencing appeared to create problems for some respondents, especially those reporting visits in answer to the preceding question. Second, the question appears to be ill-defined conceptually with regard to the fact that only dental examinations and treatments were to be included. Third, the response form is poorly specified. The interviewer was required to code the answers into categories "within the last two weeks," "more than 2 weeks to 6 months ago," etc., but the respondent was not made aware of these categories.

Question 23 read "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 gave qualified adequate answers, expressing uncertainty about their responses. It appears that a number of respondents found it difficult to make the comparison required by the question.

Question 30c was the third in a series of questions that were prefaced by "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...." Question 30c read "... taking care of your teeth or dentures?" Interviewers were allowed but not required to repeat the response choices with any of the questions in the series. Over a quarter of the respondents gave an inadequate answer to this question. Many respondents appeared not to know or had forgotten what the response alternatives for question 30c were.

The final question in Table 1 (question 35) is one for which as many as 70% of respondents had some type of problem. The question read "We've been talking about heart disease. Now we want to talk about cancer. What do you think are the warning signs or symptoms of cancer?" There were many "don't knows," requests for clarification, and inadequate answers to this question. Our hypothesis is that most of the respondent problems stemmed from a lack of clarity in the phrase "warning signs or symptoms of cancer," a phrase popular in public health but apparently not very salient to the general public.
The overall conclusion from the analysis of the behavior codings is that they provide a valuable systematic assessment of question problems. They uncover a range of problems that are not apparent with standard pretesting and debriefing procedures. A limitation of behavior codings is that they cannot fully capture the nature of question problems. The various indicators derived from behavior codings can give general pointers to the sources of problems, but more question-specific information is often needed for a full understanding of the causes of the problems with individual questions.

4. Special Probes

Several researchers (e.g., Schuman, 1966; Belson, 1981) have used special probe questions to explore respondents' understandings of survey questions. Special probe questions may be employed in pretests to encourage respondents to elaborate on their answers, to explain how they interpreted a question, or to describe any difficulties they had with a question. The use of probe questions has the potential to reveal problems that would otherwise go undetected, as for instance when respondents readily give apparently adequate answers to questions that many of them have misinterpreted. The answers to special probe questions may also identify the sources of difficulty, and hence may indicate ways in which the question may be improved.

Two restrictions on the use of special probe questions were faced in this study. First, it was felt that only a few questions could be probed, and only certain types of probes could be used, during the course of an interview without significantly altering the nature of the interview and perhaps affecting the way in which the respondent answered subsequent questions. This restriction was overcome by deferring some of the probe questions until the end of the interview. Second, it was felt that asking many special probe questions of any one respondent would be too taxing and would lengthen the interview inordinately. This restriction on the number of questions that could be probed was overcome by dividing the pretest samples into subgroups and probing different sets of questions with different subgroups.

Several different types of special probe questions were incorporated in the questionnaires for groups 2 and 4 in this study. The 100 interviews in each group were divided into three subgroups, each with a different set of special probes. This procedure enabled 26 questions to be probed in group 2 and 22 questions to be probed in group 4.

Four or five special probe questions were embedded in the questionnaire to be asked immediately following responses to the targeted questions. These probes were designed to fit in with the flow of the interview and several were of the type used routinely by interviewers (e.g., "Could you tell me more about that?"). The remaining special probes were added at the end of the interview. They were introduced by a statement that asked the respondent to assume the new role of helper in improving the questions. Respondents were asked to talk about the meanings of the questions and their experiences in answering them. Most of the special
probes used at the end of the interview were more focused than those embedded in the questionnaire. The probes were of three main types: comprehension probes, information retrieval probes, and response category selection probes. Examples of each of the three types are:

To explore respondents' understandings of the term "red meat" in a question about food eaten, the special probe read: "Would you include things like bacon, hot dogs, or lunch meats as red meat?"

"Could you tell me about any problems you might have had figuring out the answer?"

"How hard was it for you to pick an answer that describes how you really feel?"

Apart from the comprehension probes, the other special probes were not particularly successful in identifying question problems. The general probe "Could you tell me more about that?" did uncover the fact that, in answer to a question asking respondents to rate their health relative to others of their age, many respondents appeared to rate their health in absolute rather than relative terms. This kind of probe also revealed that in answering another question many respondents misinterpreted one of the two statements between which they were to choose. However, the general probes mostly failed to find problems with questions, including four questions identified as problematic by the behavior coding.

Information retrieval probes were employed for fifteen questions. These probes revealed that a few respondents had retrieval problems with several of the questions, but only one question was shown to have serious problems. Yet the behavior codes showed that ten of these questions had substantial problems, with retrieval problems appearing to be the main cause. A possible explanation for the failure of the information retrieval probes to identify retrieval problems with these questions is that the respondents did not see themselves as having such problems, even though the behavior coding suggests otherwise.

Six closed questions were probed for respondent difficulties in selecting the appropriate response category. Although these probes gave evidence of other problems, they did not reveal problems with choice of response category. It may be that respondents did not encounter such problems, or it may be that they did not understand the probes as intended.

The comprehension probes were successful in uncovering a substantial number of problems. Fifteen per cent or more of respondents were found to have problems with the meaning of the question for 12 out of the 18 questions with which comprehension probes were used. The problems generally involved misunderstandings rather than uncertainty about the question meaning. As an example, consider the question "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?" The special probe "How clear was it to you what to include as a half a day in bed?" revealed that many respondents interpreted the term to mean not getting out of bed in the
morning and staying in bed until noon or later, while others interpreted the term to mean a length of time, from 2-4 hours to 12 hours or more.

Another example relates 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 was probed intensively at the end of the interview, with respondents being asked which of a series of health professionals (chiropractors, physical therapists, nurses, etc.) should be covered by the question. There was considerable disagreement among respondents as to which health professionals should be included and which excluded. Response to another special probe on this question revealed that one-third of respondents thought that a telephone conversation for medical advice should be included, but the other two-thirds thought that it should be excluded.

The general conclusion from this study is that special probes appear to be an effective method for uncovering problems with question understanding, particularly with lack of common understanding and misinterpretation of the question. The kinds of special probe used in this study to identify other types of problems were not very effective. This finding suggests that special probes may not be particularly suitable for revealing other types of problem, but it may only indicate that different types of probe question are needed.

5. Special Training of Pretest Interviewers

Pretesting is commonly carried out by interviewers who have little or no special training for the task. After the interviewers have completed their interviews, they meet with a moderator (often an interviewing supervisor) and the researchers for a debriefing session. At the debriefing, the interviewers give comments on their experiences with the questionnaire. The limited success of standard pretesting in identifying question problems may in part be due to the lack of interviewer training in identifying problems and to the unsystematic way in which they are asked to report their experiences in the debriefing session.

In this study an attempt was made to improve on standard pretesting by modifying the usual procedures in two main ways. These modifications were applied to groups 3 and 5. One modification was to put the interviewers through a special training program. This program, which lasted about five hours, addressed the question as a source of problems for interviewers. The training explained the goal of identifying problems with questions, and provided practice with problem identification. It focused on questions that are difficult to read, problems of definition or understanding, questions that pose problems for respondents in knowing what the answer is, and questions that are difficult to answer in the form required. As part of the training, interviewers listened to tape-recorded interviews and discussed the problems that occurred.
A second modification was designed to yield a more systematic evaluation of the survey questions. For this purpose, the interviewers reviewed the completed questionnaires after each interview and made notes of problems on their individual master questionnaires. After completing all the interviews assigned, each interviewer then filled in a standardized rating form that evaluated each question on four dimensions: whether the question was difficult to read as written; whether the question posed problems for the respondent in understanding its meaning; whether the question posed problems for the respondent in knowing what the answer was, either because of recall difficulties or because the question asked for an opinion on an issue that the respondent had not previously considered; and, whether the question posed difficulties to the respondent in producing an answer in the form the question required. Each question was evaluated on these four dimensions using a three-point rating scale of no problem evident, possible problem evident, or definite problem evident.

The interviewers for groups 3 and 5 were debriefed by an experienced interviewer supervisor who had the interviewers discuss each question. Although the discussion was similar to that in usual debriefing sessions, it differed in that interviewers were asked to present their ratings as part of their discussion, and efforts were made to understand differences between different interviewers' ratings. As a result, the debriefing session took much longer than usual to complete. After the debriefing session, the interviewers filled out their rating forms again to reflect any changes in their judgments based on the debriefing discussion.

Analysis of the results for the specially trained interviewers is still underway. The overall impression is that the interviewers' ratings yielded similar results in terms of types of problems identified with individual questions to those obtained by standard pretesting and debriefing. A possible exception was a greater number of questions that were classified as having reading problems by the interviewers' ratings, an exception that may be explained by the attention given to reading problems in the special training program.

In a standard debriefing session many of the interviewers may not contribute to the discussion on any one question. It is therefore difficult to assess the prevalence of particular problems. A potentially valuable feature of the standardized rating form is that it collects interviewer comments and impressions in a systematic way which makes it convenient for researchers to use in evaluating each question.

6. Discussion

Respondents in group 1 of the study were interviewed using standard pretest procedures and the interviewers were debriefed at a standard debriefing session. The respondents in group 2 were asked a number of special probes; apart from this, they were also interviewed using standard pretest procedures and the interviewers were debriefed at a standard debriefing session. Content analyses of the tape recordings of these debriefing sessions revealed that the standard pretest procedures failed to identify an appreciable number of the question problems identified by the
other methods, and that they failed to provide adequate estimates of the prevalence of the problems reported. These findings support the case for improved pretesting procedures.

One way to improve questions is to identify potential problems prior to field testing. The use of a laboratory setting to study questions in the design phase has recently been introduced at the U.S. National Center for Health Statistics, the Bureau of Labor Statistics, and the Bureau of the Census. Descriptions of this approach are given by Royston, et al. (1986), Royston and Bercini (1987), and Lessler and Tourangeau (1986). While laboratory testing may provide in-depth understanding of respondents' comprehension of survey questions and the cognitive processes they employ in arriving at their answers, it does not replace field testing. Rather, the two serve complementary functions.

The research reported here suggests several methods for improving the standard procedures for field testing survey questionnaires: the use of behavior coding; the use of special probes; training interviewers for pretesting; training moderators in how to run debriefing sessions; and having the interviewers complete standardized rating forms on the survey questions. The issue to be resolved is how to put together a combination of these methods that will provide an effective and efficient field pretesting procedure.

The behavior coding proved its value in this study. The analysis of behaviors produced a set of problem indicators that brought to light problems not identified by other methods. The indicators provided quantitative information on the prevalence of the problems. Some form of behavior coding is a valuable addition to standard pretest procedures. The coding scheme used in this experimental study was a detailed one, classifying each separate component of a respondent's behavior. The analysis of these detailed codes suggests that a simpler coding scheme would suffice; since the indicators given in Table 1 capture the important aspects of a respondent's behavior, the coding scheme can be reduced to only behaviors that are needed for these indicators. A simplified behavior coding scheme could be easily applied to tape-recorded interviews or while the interviews are being conducted. It would not be difficult to prepare a table like Table 1 for all the survey questions in time for use at the debriefing session. The interviewers could then be asked at the debriefing to interpret the indicators and give their views on the source of the problems that the indicators had identified.

The analysis shows that special probes are also useful in identifying particular kinds of question problems. They are most effective when the researcher wants to explore respondents' understandings of a concept in a question. Carefully designed special probes on understandings of concepts may identify problems that would not be apparent with any of the other procedures. Our experience suggests that designing effective diagnostic probes is difficult and requires considerable work and skill.

Although analysis of results for interviewersa given special pretest training is still in progress and no firm conclusions can be reached, it
seems likely that some training to sensitize interviewers to the types of problems that can occur with survey questions may be useful. In addition, it may be helpful to give the moderator of the debriefing session instruction in how to run the session in a way that will allow problems to surface and will secure participation of all the interviewers. The use of standardized rating forms by the interviewers to record their views on the problems with each question provides a systematic accounting of their views, and the ratings can provide a basis for discussion at the debriefing session.

In summary, a possible scheme for pretesting survey questionnaires might be along the following lines. Prior to conducting their interviews, pretest interviewers are given a brief training to sensitize them to the types of problems that might arise. The questionnaire may contain some carefully prepared special probes to explore respondents’ understandings of certain questions. The interviews are behavior coded immediately, using a simple behavior-coding scheme, and the behavior code indicators are tabulated in time for use in the debriefing session. The interviewers may, or may not, be asked to complete a standardized rating form. If interviewers’ ratings are available, the moderator for the debriefing session is instructed to use them along with the behavior code indicators in a systematic evaluation of each question. While we believe that a pretesting scheme of this type would improve on present practices, further research is needed to refine and integrate the various pretest procedures.

In conclusion it should be noted that our ultimate objective is to improve the survey questionnaire. Effective pretesting will identify many of the question problems that exist and provide information on the causes of the problems. When problems are identified, the questions can be modified in an attempt to improve them. In this study we revised the questions in the light of the insights gained from a first pretest and administered the revised questions in a second pretest. Generally, the revisions were successful in eliminating the question problems. In some cases, the revision removed the original problems but introduced new ones. In a number of cases involving difficult concepts, the revisions reduced the level of the problem but a significant problem still remained. It is often impossible to deal with severe conceptual problems by simply changing the wording of a question. Rather, to eliminate such problems may require significantly different questioning strategies, and in extreme cases it may even be impossible to design meaningful questions. Despite the difficulties involved in rewording questions in some complex cases, overall we are convinced that the methods proposed here can significantly improve survey questions and lead to marked improvement in the quality of survey data.

Acknowledgement

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


SUMMARY

This paper reports the results of a study conducted to test the effectiveness of some new techniques for improving standard pretesting procedures. The new techniques are: the use of behavior coding; the use of diagnostic probe questions; the training of pretest interviewers in identifying question problems; and the completion of standardized rating scales by the interviewers to report their evaluations of the questions. The conclusion reached from the study is that the inclusion of a combination of these techniques in standard pretesting would improve the identification of question problems and of the causes of the problems.

RÉSUMÉ

On présente ici les résultats d'une étude faite pour vérifier l'efficacité de quelques nouvelles méthodes pour améliorer les techniques utilisées dans le "pretesting." Les nouvelles méthodes sont: l'utilisation de "behavior coding"; l'utilisation de questions pour faire des relances; l'entraînement des enquêteurs en vue d'identifier les questions qui posent des problèmes; et l'utilisation de "standard rating scales" par les enquêteurs mêmes pour indiquer leurs jugements quant au valeur des questions posées aux interviewés. L'utilisation d'une combinaison de ces techniques faciliterait l'identification de questions difficiles dans les "standard pretests" aussi bien que les causes de difficultés.