BEFORE THE PRETEST: QUESTION DEVELOPMENT STRATEGIES

Steven Blixt and Jennifer Dykema, The University of Michigan
Steven Blixt, Institute for Social Research, 426 Thompson, Ann Arbor, MI 48106-1248

KEY WORDS: Questionnaire design, pretests, behavior coding, intensive interviewing

While the importance of pretesting survey instruments is well-accepted, conventional methods tend to be unsystematic and relatively ineffective (Cannell et. al, 1989). Some researchers have attempted to address weaknesses in the pretest process by using other question development strategies. One such strategy is to do intensive interviewing, usually as part of a "pre-pretest." In intensive interviewing, special techniques are used to uncover information about question comprehension and other response difficulties that may not be elicited without extensive probing. A disadvantage of this technique is that it produces data that is sometimes difficult to analyze objectively. A second strategy for improving the pretest is to conduct behavior coding of interviewer-respondent interactions. This coding systematically identifies potential question problems, but does not always diagnose the sources of the problems it finds.

In our paper, we report on a new method that merges the complementary strengths of intensive interviewing and behavior coding. By applying behavior coding techniques to the analysis of intensive interview data, **systematic intensive interviewing** provides qualitative as well as quantitative information for use in the early stages of questionnaire design. We briefly summarize these two strategies and discuss how they were integrated as part of a larger research project to develop and test questions for inclusion in the 1995 redesign of the National Health Interview Survey (NHIS).[1]

The First Strategy: Intensive Interviewing

Intensive or cognitive interviews are typically conducted in a lab setting by interviewers with training in cognitive or clinical interviewing methods. Using procedures such as verbal report techniques (think-alouds), paraphrasing, and memory cues, intensive interviews investigate how successfully respondents work through the information-processing and decision-making steps involved in answering survey questions (Jobe & Mingay, 1989; Cannell, Miller, & Oksenberg, 1981; Tourangeau, 1984). Some of these techniques are designed to expose hidden comprehension problems and information retrieval difficulty. Others increase understanding of the judgment strategies that ultimately determine an answer. Another focus of intensive interviews is motivation and affect. Motivation determines how thoroughly a person accepts the respondent role: how diligently he or she attempts to comprehend what is wanted; how much effort is directed to searching memory; and finally how much potential embarrassment the respondent will tolerate to report accurately.

While their contributions to question development are becoming well-recognized, intensive interviews do have certain drawbacks. Differences between lab and field conditions can be substantial and may influence findings. In intensive interviews, probes disrupt the interview flow; interviews are conducted without the distractions found in a household setting; the respondents are usually highly motivated and sample sizes small and often unrepresentative. Perhaps most significant is the absence of rigorous standards for analyzing the data. How are problems identified and prioritized? What effort is made to systematize this process? The literature shows analysis procedures for intensive interviews are often subjective, suggesting these results remain more impressionistic than scientific (Aday & Kasper, 1989).

The Second Strategy: Behavior Coding

Behavior coding enhances the questionnaire development process by quantifying the problems that occur in interviewer-respondent interactions. Codes are assigned to specific respondent behaviors such as interrupting question reading, pausing before answering, qualifying an answer with "about" or "I guess;" requesting clarification; providing answers that do not match the response options; and saying "don't know" or refusing to answer. Frequencies of these codes are computed on a question by question basis, and sometimes a summary measure is used to indicate the proportion of respondents who exhibit any of these behaviors for a given question. By comparing results across questions, researchers can identify those items that require closer examination, revision, and perhaps elimination. If another pretest is conducted, the distributions from the first pretest can serve as a benchmark for judging whether rewriting a question has reduced the levels of problem indicators (Fowler, 1992).

While behavior coding helps to systematize the identification of question problems, it too has limitations. The researcher must still infer the source of a problem in order to correct it. Knowing there is a high percentage of requests for clarification at a particular question is not useful unless something is known about the nature of the request. Further, behavior coding does not help detect when respondents mistakenly believe they understand the question objective, or are unwilling to admit they really do not understand or don't know an answer.

The Solution: An Integrated Approach

Recognizing the strengths and weaknesses of the questionnaire development methods just described, we decided on an integrated approach to test items under consideration for the 1995 redesign of the NHIS. Working in conjunction with National Center for Health Statistics, we designed an intensive interview protocol in which draft survey questions were followed by structured probes. Audiotapes of the interviews were then analyzed using a behavior coding scheme modified to permit quantification of information from intensive probing. In this way, our approach built upon the strengths of intensive interviewing and behavior coding in order to develop a pretesting strategy more systematic and informative than either technique affords alone. This strategy produced results that enabled us to make informed decisions about retaining, revising, or eliminating items for subsequent testing in a traditional pretest.

Methods

Phase I: Intensive Interviewing

A total of 87 interviews were conducted in lab or office settings in Ann Arbor and Detroit during two separate rounds of interviewing. We recruited respondents through announcements in local newspapers offering $20 to those wishing to take part in a study on health-related issues. Attempts to recruit across a range of demographic characteristics were successful. Topics of the interview included quality of life, as well as traditional areas such as physical health status and role functioning. Questions were taken from the current NHIS, other existing instruments, or were drafted by the research staff. Interviews averaged an hour in length and were audiotaped.

Five staff members trained in cognitive interviewing techniques conducted the interviews. In introducing the survey, interviewers emphasized to subjects their dual role as both respondents to the survey questions and as informants regarding their cognitive processing and affective reaction to the questions. Interviewers were told to ask survey questions exactly as written so that we could do behavior coding using a controlled stimulus. Numerous structured probes were written into the questionnaire based on our hypotheses about problems a respondent might experience in answering a question. Table 1 gives a few examples of these probes. Interviewers did additional probing to illuminate negative affect, inconsistent answers, or suspected comprehension problems. Weekly follow-up training sessions were held in which interviewers received further instruction or feedback regarding their performance.

Phase 2: Coding

Audiotapes of the interviews were coded in four ways. First, respondents' answers to the survey questions were tabulated. Second, responses to most of the scripted probes were coded using simple schemes. For instance, for a probe asking respondents what they were thinking when they rated their "overall health," answers were coded into eight substantive categories including health behavior, physical health status, and mental/emotional health status. This type of coding was used to establish the frame of reference the respondent used in answering the survey question, as well as to identify definition and other comprehension problems with particular items. Third, standard respondent behavior coding was done to analyze the initial reaction of respondents to the main survey questions. Brief descriptions of these codes are given in Table 2. Finally, we developed new behavior codes suited to the kinds of probes used in the intensive interviews.

Coders listened to the entire interaction at a survey question, from the initial response through the follow-up probes. Table 3 shows the codes used to evaluate this interaction and gives an illustration of each one. The first intensive interview codes concern displays of negative affect, either toward the survey question or the response options. Specifically, these included comments such as "that's a dumb question" or "I think some people may interpret that word differently." An African American respondent, for instance, explained she would never choose the answer category "fair" to describe her health because she believes that word is insulting to African Americans.

Table 1: Examples of Probe Techniques

Type of Probe	Survey Question	Probe
DEFINITION	How often do you plan or select meals with nutrition in mind?	When I use the word "nutrition" what does that mean to you?
CONCURRENT THINK-ALOUD	During the past month, how much of the time did you feel self-confident?	Sometimes learning how a person goes about figuring out an answer helps us write more sensible questions. After I read the question, I want you to tell me everything that's going through your mind--in other words think out loud...
FRAME OF REFERENCE	How true is the following statement for you? I seem to get sick a little easier than other people.	When I said "other people," what sorts of people did you think of?
MOTIVATION/ AFFECTIVE	During the past two weeks on how many days did you drink alcoholic beverages such as beer, wine or liquor?	Do you find it embarrassing to talk about how often you drink?

Table 2: Standard Respondent Behavior Coding Scheme

Code	Code Description
Int	Interrupts question-reading with an answer.
Pau	Pauses/hesitates noticeably before answering.
Qlf	Gives qualified, codeable answer. *I guess*, *around*, and *about* are examples of qualifiers.
Clf	Requests clarification: asks for repeat of the question, part of the question, or clarification of a specific term, phrase, or concept (e.g. "Did you say past 30 days?", "What do you mean by 'impairment?'").
Unc	Gives an uncodeable answer. An uncodeable answer is any response to a close-ended question that does not match the options provided. (e.g. When the respondent is asked to indicate on 0 to 10 scale how much pain interferes with his or her activities, the respondent says "all the time.").

Table 3: Intensive Interview Respondent Behavior Coding Scheme

Code	Code Description	Code Illustration
CrtQ	Criticizes survey question; expresses negative affect, suggests change(s) to survey question for self or others.	When asked about strenuous activities such as jogging, running, playing handball, tennis, or swimming, the respondent said these examples were "very privileged exercise things, but that's not what strenuous activity equates to for most people."
CrtR	Criticizes response options; expresses negative affect, suggests change(s) to response options for self or others.	The respondent felt the response options "a good bit of the time" and "most of the time" were "too close."
CmpQ	Indicates comprehension difficulty with a term, phrase, or concept in response to intensive probing.	When asked what foods came to mind when he thought of "red meat," respondent said "chicken."
CmpR	Indicates comprehension difficulty with response options in response to intensive probing.	The respondent said (referring to -5 to +5 number scale) "I think the scale is reversed, +5 should be the mild pain and -5 should be the worst pain because '+'is a positive way of looking at things."

In addition to looking for indications of affect, coders also used information gathered from follow-up probing to determine whether there was evidence the respondent might not have understood the question or response options. To make this assessment, coders evaluated the respondent's understanding of terms and concepts against the meaning as intended by the question writer. Such criteria was necessary since respondents sometimes offer logical interpretations of questions that nevertheless differ from what the question writer wants to convey. For instance, at an item about the respondent's physical health status (referred to in the question simply as "health"), probing revealed that half of all respondents had answered the question partly or entirely on the basis of health behavior.

In addition to assigning codes, coders wrote notes regarding the nature of response and comprehension problems. These notes were made whenever codes for clarification requests (Clf), uncodeable answers (Unc), or any intensive interview codes were assigned. The notes assist question writers in diagnosing the source and the seriousness of the problem.

Three experienced interviewers coded all of the interviews. They were trained on the criteria to use in assigning codes as well as the objectives of individual questions. During training coders coded as many practice interviews as was necessary in order to ensure they were assigning codes correctly. Once production coding began, 12 of the 87 cases were independently coded by two coders and the kappa statistic was used to assess inter-rater agreement. The reliability scores of all standard and intensive interview codes fell into the range of fair to good agreement, .40 to .75 as described by Fleiss (1981). Mean reliability of the standard behavior codes was .65 with the least reliable standard code being qualified answers (.45) and the most reliable, uncodeable answers (.75). Among the intensive interview codes, the overall mean was .63 with question comprehension being the least reliable (.49) and negative affect toward the response options the most reliable (.71). The kappa statistic was not computed for codes that were assigned very infrequently.

Results

Through analysis of coding data, questions were classified into three general categories: items that appeared to work well, those that required modification, and finally, some that had problems so major they needed to be eliminated from further NHIS consideration. We identified the worst questions by reviewing the standard and intensive interview behavior coding results. We examined each question individually and made judgements about how serious the problem was and what trade-offs would be involved in revision. Variance in the distribution of survey question responses and the answers to qualitative probes were used to supplement and help interpret the information from behavior coding. Our analysis thus took into account both the qualitative and quantitative information from the intensive interview.

In Table 4, two questions are presented to illustrate how this process worked. First, the distributions to the standard behavior coding were examined. These codes were assigned based on behaviors exhibited by the respondent in answering the survey question only. Principal emphasis was placed on requests for clarification (Clf) and uncodeable answers (Unc). Levels of 15% or more for either of these behaviors were considered high. Emphasis was also placed on the summary measure (Sum) which gives the percentage of respondents who were assigned at least one standard code. Percentages from the intensive interview behavior codes are also given.

A. *On about how many of the past 7 days did you eat foods that are high in fiber, like whole grains, raw fruits, and raw vegetables?*

At question A, over half of all respondents exhibited one or more problem indicators (the Sum column under 'Standard Behavior Codes'). A large percentage gave qualified answers such as "about 3" or "around 2," others requested clarification of the term "high fiber" asking "does that include breads?," and "can they (vegetables) be steamed?" Still others provided uncodeable answers like "twice," or "not enough." After answering the survey question, respondents were asked what "high fiber" meant to them. About a quarter gave responses that indicated possible comprehension problems. Several admitted they did know what the term meant or could not define it more specifically than foods that are "high in fiber;" others gave definitions that were partially or completely incorrect. One respondent described it as "the foods you consume," another said it meant "oats, tomatoes and vitamin C."

B. *How true is the following statement for you? I seem to get sick a little easier than other people. Would you say that statement is definitely true, mostly true, mostly false or definitely false?*

Examining results for question B, relatively few

Table 4: Examples of Standard and Intensive Interview Behavior Coding Results (in Percentages)

Standard								Intensive Interview Codes					All
Q	n=	Int	Pau	Qlf	Cla	Unc	Sum	CrtQ	CrtR	CmpQ	CmpR	Sum	Sum
Fiber	42	2	7	26	19	14	55	0	0	24	0	24	69
Sick	33	3	3	3	12	21	27	9	48	6	36	70	47

NOTE: Other standard behavior codes were for don't knows and refusals. The other intensive interview code was for instances when respondents changed their answer to the original survey question after being probed. These were infrequently assigned and are not included here.

Key to Codes

Sum = Summary measure
Int = Interrupts
Pau = Pauses
Qlf = Gives qualified response
Cla = Requests clarification
Unc = Gives uncodeable response

CrtQ = Criticizes survey question
CrtR = Criticizes response options
CmpQ = Indicates comprehension difficulty with survey question
CmpR = Indicates comprehension difficulty with response options

respondents exhibited one or more problem indicators. Most of the coded behaviors involved requests for clarification such as "what do you mean by 'other people'" or uncodeable answers, for instance answering "false" without specifying "mostly" or "definitely." This standard behavior coding summary score is below the average for all questions in the interview (mean=34%), which means this question would not stand out as problem in a pretest using standard behavior coding alone. However, nearly three-quarters of the respondents expressed negative affect or indicated some comprehension difficulty upon being probed. This figure is three times the mean value for the intensive interview behavior coding summary score. Reaction was particularly strong regarding use of the word "mostly." Some respondents felt it was too much like "definitely," others said it made no sense to them in this context saying "either it's true or it's not." This question thus demonstrates the way in which behavior coding and intensive interviewing can operate together to uncover and objectively assess question problems that might otherwise be missed.

After examining the behavior coding results, decisions were made whether to retain or revise questions for the next stage of item development, a pilot survey administered by regular Survey Research Center field interviewers. Question A was rewritten to remove the word "about," which it was felt might be encouraging fuzzy, qualified answers, and omit the phrase "high fiber" which had been poorly understood. Problems with question B were judged to be too serious and widespread to be easily solved, therefore the item was eliminated.

Discussion

A formal pretest should be a time for making refinements, smoothing out the interview flow, determining section lengths, and spotting technical problems with the questionnaire such as incorrect skip patterns, formatting and typographical errors. Expecting pretest interviewers to find and objectively report on all of the important problems in a questionnaire seems unrealistic. Perhaps equally unrealistic is expecting that all major question problems can be solved in the period between the pretest and the start of production interviewing. We believe that question testing must therefore begin before the formal pretest.

Overall, the questionnaire development strategy described here appears to be an effective procedure for evaluating questions. By merging intensive interviewing with behavior coding, we have systematized the analysis of intensive interview data so that the results can be used in a more rigorous way. In addition, we have improved behavior coding as a pretest strategy by using cognitive techniques to flesh out the sources of question problems that would otherwise be left to inference, and by discovering problems that are not revealed in typical pretests.

Researchers intending to use this method are strongly urged to determine the objective of each question before they begin. This step seems basic but is easy to overlook. If the investigator lacks clear understanding of an item's intent, it becomes difficult to test or to diagnose problems. Scripted probes should be based on some hypothesis the researcher has about problems with a particular question because poorly written or extraneous probes only waste valuable interviewing and coding time. During our analysis, we occasionally had to sort out whether a respondent's comprehension problem or affective response were in reaction to the survey question or to the scripted probe.

As with any new methodology, systematic intensive interviewing requires further developmental work. This work might be focussed on the interview, the coding process, and on the way in which the two techniques are integrated. For example, certain cognitive techniques work better than others in uncovering problems with survey questions. More research is necessary to determine the most effective techniques. This method also requires well-trained and skilled interviewers. Systematic intensive interviewing challenges interviewers because it requires them to continually switch from reading survey questions in an exact, standardized manner to doing follow-up probing in a conversational tone using their own or written probes. Despite our training and continuous monitoring, we were unable to eliminate all important differences in interviewer probing performance. In summary, the interviewing process needs to become less specialized so that it can be implemented quickly and cost effectively.

Coding interviewer question-reading behavior may be a useful addition to this methodology. Such coding is usually done in pretests that are behavior coded, but was not done in this study because of time constraints and what we perceived as differences between lab and regular field interviewers. In retrospect, evaluating how well interviewers ask a question in the lab may shed light on problems interviewers will have in the field. Regarding other coding matters, the intensive interview codes developed for this process focus on affect and comprehension. There may be other dimensions of respondent behavior that particular researchers want to code. More basic, however, is determining data quality. To what extent do comprehension or motivation problems lead to poor reporting? Currently, evidence of this relationship is not well-established. A validation study is necessary in order to understand the ultimate implications of behavior coding results for response accuracy.

[1] This research was supported by a Cooperative Agreement with The National Center for Health Statistics through the Association of Schools of Public Health. The views presented here do not necessarily represent the views of these agencies.

References

Aday, L., & Kasper, J.D. (1989). Strategies for evaluating questions. Proceedings of the Health Survey Research Methods, National Center for Health Services Research and Health Care Technology Assessment, 47-50.

Cannell, C.F., Miller, P.V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt, (Ed.), Sociological Methodology. San Francisco: Jossey-Bass.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., & Fowler, F.J. (1989). New techniques for pretesting survey questions. Research Report. Survey Research Center, The University of Michigan.

Fowler, F.J. (1989). Coding behavior in pretests to identify unclear questions. Proceedings of the Health Survey Research Methods, National Center for Health Services Research and Health Care Technology Assessment, 9-12.

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

Jobe, J.B., & Mingay, D.J. (1989). Cognitive research improves questionnaires. American Journal of Public Health, 79(8), 1053-1055.

Tourangeau, R. (1984). Cognitive science and survey methods. In T.B. Jabine, M.B. Straf, J.M. Tanur and R. Tourangeau (Eds). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, (73-100). Washington, DC: National Academy Press.