

# THE USE OF NOVEL PRETESTING METHODS IN THE DEVELOPMENT OF SURVEY QUESTIONNAIRES

Gordon B. Willis, Deborah Trunzo, and Barbara J. Stussman  
National Center for Health Statistics

Gordon Willis, Rm. 915, 6525 Belcrest Rd., Hyattsville, MD 20782

**Key Words:** Survey pretesting, cognitive methods

Over the past several years, the pretesting of large Federal surveys, such as the NCHS National Health Interview Survey, has undergone a set of evolutionary changes. The most noteworthy of these has been the inception of the use of cognitive laboratory techniques to uncover sources of response error that were not otherwise detectable. At NCHS, the Questionnaire Design Research Laboratory (QDRL) has operated for this purpose since 1986, and similar efforts have been undertaken at the U.S. Bureau of Labor Statistics (Dippo, 1989), at the Census Bureau (Campanelli, Rothgeb, and Martin, 1989), and at Statistics Canada (Gower and Dibbs, 1989). An overview of this work is given by Jobe and Mingay (1991). We have found the addition of these laboratory techniques to be extremely useful (Willis, Royston, and Bercini, 1991), but also limited in several key ways:

a) For laboratory interviews, individuals are self-selected into participation, usually by answering advertisements and flyers. Therefore, they may differ in important ways from the population to be surveyed. For example, Stein (1992) has determined that laboratory volunteers tend to be higher in education than are members of the general population.

b) The interviewers in the laboratory are survey researchers, and not professional Census Bureau interviewers. There are important differences between these two types of individuals, related to their interactions with survey respondents.

c) The context of the laboratory interview differs greatly from that of the household, where an interviewer may simply arrive at the door, interrupting the respondent's usual activities. Further, household respondents are not paid, and may therefore be substantially less motivated to answer carefully.

One might argue that the traditional field pretest has none of the shortcomings of laboratory testing, but this method too is limited in important ways:

a) The field pretest is, in actuality, more often a "dress-rehearsal" of survey procedures than a focused test of the questionnaire instrument. Subtle

questionnaire-based problems are often undetected, or are given lower priority than administrative problems.

b) Even if major questionnaire-based changes were the focus of the field pretest, the pretest generally occurs so late in the developmental process that there is insufficient time for large-scale modification.

In order to address the concerns outlined above, we have tested a set of two additional pretesting activities for the development of survey questionnaires. These activities are conducted after an initial phase of laboratory interviewing has been completed, but before the regular large-scale field pretest:

1. A laboratory-based review of the draft supplements by NHIS interviewers;
2. Pilot household pretesting by NHIS interviewers.

Evaluation of the efficacy of these activities will be the focus of the remainder of this report.

Questionnaire. The particular questionnaire serving as the focus of the testing to be described was the Cancer Control and Epidemiology Supplement to the 1992 National Health Interview Survey (NHIS), an interviewer-administered survey of 50,000 households conducted annually by NCHS. The Cancer Supplement consisted of two split-half sample person questionnaires (labelled Cancer Control, and Epidemiology), with a small degree of overlap between the two versions. The questionnaire was sponsored by the National Cancer Institute of the National Institutes of Health.

## I. Laboratory-Based Interviewer Review

The objectives of this activity were:

- 1) to obtain experienced interviewers' comments on the questionnaire content, mechanics, and wording before the questionnaire underwent extensive field testing, especially related to the cognitive-based problems interviewers anticipate;
- 2) to demonstrate to all parties involved in the questionnaire design process, i.e. sponsors, and NCHS staff, how the draft questionnaire functions

when administered by an NHIS interviewer.

### Procedure

The day-long interviewer review involved four experienced Census Bureau interviewers. The week before coming to the QDRL, they were sent a copy of the questionnaire and a cognitive rating form. Because the supplement contained hundreds of questions, interviewers were asked only to rate those items that they thought would have "a problem of some kind." There were four rating categories: **reading problems** (whether the question would be difficult for interviewers to read), **understanding problems** (whether the interviewer thought the question would be difficult for respondents to understand), **knowledge/recall problems** (whether the interviewer thought respondents would have inadequate knowledge or be unable to recall the desired information), and **other problems** (an open-ended category where interviewers could write in their comments). These initial ratings were collected when interviewers arrived at NCHS.

Once at NCHS, each interviewer participated in two interviews. For the first interview, the interviewers were divided into pairs, with one interviewing the other. For the second interview, each interviewer administered the questionnaire to an "outside" respondent (several NCHS employees not familiar with the questionnaire, and a survey professional from the Census Bureau). The interviews were observed by NCHS staff and staff of the sponsoring agency. After each interview, interviewers were given the opportunity to make informal written notes on difficulties they had encountered. After the second interview, they completed the rating form again. These activities took the entire morning. In the afternoon, the four interviewers were debriefed as a group, with all observers attending.

### Evaluation of the interviewer-review activities

**Pre and post-interview rating activity:** This aspect of the interviewer review was not a success in and of itself. Most importantly, the cognitive rating categories were not meaningful to the interviewers; virtually all of their ratings were made in the "Other Comments" column. Interviewers did not speculate about the cognitive problems respondents might have, but were adept at pointing out skip errors, confusing formatting and instructions, and other structural deficiencies that made it difficult for them to find their way through

the questionnaire document.

**Interview exercises and debriefing:** The two practice interviews conducted by each interviewer formed the substance of the interviewer review. These interviews proved sufficient to expose numerous questionnaire flaws, particularly those aspects of the questionnaire that were impediments to the flow of the interview. Interviewers pointed out many mechanical flaws, overly verbose and awkwardly worded questions, and seemingly redundant questions.

The practical utility of this exercise was greatly enhanced by the presence of members of the sponsoring agency as observers. Seeing the interviewers struggle with some of the most awkward and burdensome sections of the questionnaire motivated the sponsors to make some needed changes. Although the need for some of these changes had already been identified during lab testing, seeing the entire questionnaire administered in a more "real-world" format was quite convincing.

On the basis of our experience with the interviewer review process, we would recommend this as a useful pretesting activity, with one major cautionary qualification: The questionnaire must already exist in a developmental state sufficient to make meaningful a critique of layout, organization, and wording. There is little point in conducting an interviewer review with an embryonic questionnaire, because significant changes in basic approach and content subsequent to this activity would render the exercise fairly useless.

## II. Pilot Household Testing

The second pretesting activity had two major objectives:

- 1) to conduct limited field testing at a time when significant changes can be made if they are found to be necessary;
- 2) to obtain more information about respondent reaction to questions than can be accomplished in the traditional field test setting or in the cognitive laboratory.

### Procedure

Approximately one month after the interviewer rating exercise described above (and after moderate changes had been made to the questionnaires, based partly on the results of that exercise), but prior to the normally scheduled field pretest, three experienced NHIS interviewers were recruited by the Census Bureau to work at NCHS for five days

to conduct interviews in local neighborhoods with the draft Cancer Risk Factor questionnaires. Our goal was to complete 40-50 interviews, evenly distributed between the "Epidemiology" and the "Cancer Control" versions of the supplement. Interviewers were sent a copy of the questionnaire and a general description of the upcoming week's activities, but they were not asked to complete any home-study exercises.

Interviews consisted of a short series of screener questions, followed by the supplement questions related to cancer. The screener obtained the household roster, recording sex, smoking and employment status for each household member. Interviewers were instructed to select sample person respondents according to an algorithm, based on these characteristics, which ensured that demographic subdomains were fairly evenly represented.

Interviewing took place in local household segments selected by the Census Bureau. Particular housing units were not pre-selected, and no detailed household listing was performed; interviewers simply travelled from door to door until they found households containing eligible respondents willing to be interviewed.

The schedule for the week of activities consisted of a half day of training, three-and-a-half days of interviewing, and a half-day debriefing. All household interviews were observed by a staff member of either the sponsoring agency or NCHS.

Although reasonable efforts were made to adhere to standard NHIS formatting conventions, the questionnaires did not take the appearance of the usual NHIS pretest documents. However, the interviewers were able to adapt to the unusual questionnaires with little difficulty.

In order to gain experience with another new mode of question evaluation, the interviewer-respondent interactions in each interview were systematically recorded by the observer, using the behavior coding system developed by Cannell and his colleagues (Cannell and Robison, 1971; Fowler, 1989). In brief, behavior coding consists of assigning one or more standardized codes to individual questions based on particular respondent and/or interviewer behaviors. These behaviors, such as: "interviewer did not read question as worded", "respondent asked for clarification", and "respondent did not give a codable response", serve as quantifiable indicators of problematic questions. One goal of the coding activity was to determine the feasibility of "live" behavior coding, where the

observer performs the coding during the course of the interview, rather than working later from recordings, as has been done in previous studies. A second goal was to determine how coding results would compare to the subjective information to be gathered in the debriefing. Willis (1991) has documented the procedures used in detail.

All observers were given a two-hour training session on the coding method. Forty-three of the 49 completed interviews were coded and 47 of the interviews were audio-taped (the purpose of tape-recording was simply to determine the number of respondents who would consent to this request, in anticipation that analysis from recordings might be conducted in future pretests; only two respondents refused to allow the recording).

We considered, but rejected, the addition of a probing procedure to the household interviews that would serve as a field equivalent to our current cognitive laboratory practice. We felt that a hybrid laboratory/household interview would not be as useful as either the laboratory interviews that had already taken place, or the normal household interview, in which the respondent is simply asked to answer the questions.

#### Evaluation of pilot household interviewing

The testing provided two forms of feedback: interviewers' comments from the debriefing session, and the behavior coding results.

a) Interviewer debriefing results. The half-day interviewer debriefing was conducted in the traditional style, in which comments on each section of the supplement were solicited by a discussion leader. Because only three interviewers participated in the testing, there was ample opportunity for each to report their experiences.

As a general rule, interviewers' comments focused on three kinds of questionnaire problems:

- 1) Interviewers were very critical of questions they considered too wordy or awkwardly phrased, and questions that contained long lists in either the stem or the answer categories.
- 2) Interviewers were adept at identifying questions that frequently elicited qualified or otherwise uncodable responses.
- 3) As expected, the interviewers pointed out inefficient or incorrect skip patterns, awkward or confusing formatting, and related difficulties.

While some of these problems had been evident in lab testing, they were not the focus of the cognitive interviewing process, and in some cases did not seem as acute in lab testing.

b) **Behavior coding results.** Willis (1991) has evaluated the behavior coding component of the informal household testing in detail. In brief, problems identified through the coding process coincided with problems detected through the informal debriefing process. The coding data therefore added little additional information. However, the coding results, which were available a week after the pretest, were very useful in persuading sponsors and others that some of the questionnaire flaws were substantial.

The coding activity had several side-benefits. We learned that most respondents do not object to having the interview audio-taped. Because observers were responsible for the taping, interviewers were not burdened with additional equipment or chores. However, coding interviews "live" may not be an efficient or feasible procedure in most cases (every interview has to be observed by a trained coder). Instead, it may be more useful to rely on subsequent coding of the recordings of interviews, by experienced coders. The practice of coding from recordings would also serve to address another limitation of the approach we used: our coders were not highly trained, and inter-coder reliability was therefore undoubtedly low.

### **Conclusions**

In sum, all of the stated objectives for the informal household testing were met, in addition to some unanticipated ones. It was clearly demonstrated that pilot household testing using professional Census Bureau interviewers can yield a great deal of valuable information about the draft questionnaire, and that interviewers are willing and able to adapt to unconventional pretest routines.

### **Recommendations for Questionnaire Pretesting**

**1. Questionnaire development and testing activities for large surveys should be seen as a continuous, integrated process, rather than as a set of independent activities.**

Given the time available for the development of a given questionnaire, a customized testing plan should be developed, from initial concept development through laboratory and field testing. We feel strongly that the selection of pretesting methods, such as cognitive interviewing, expert review, behavior coding, and small-scale pretests, is not an "either-or" situation in which the different techniques are tested to determine which is "better" than the others. Rather, we view each technique as having particular value at certain stages of the

questionnaire development process.

**2. To the extent possible, questions that have undergone major revision after the field pretest, or that have been added after the field test, should be tested in the cognitive laboratory before being finalized for the main survey.**

When extensive testing efforts have been devoted to previous questionnaire versions, it is incongruous that the questionnaires should go into the main survey containing revisions that have not been tested at all. Some effort can be made to test these new questions, even if on a minor scale.

**3. New evaluative pretesting activities should not automatically include features that rely on the explicit cognitive modelling of the survey response process.**

There is a strong temptation for those of us who are associated with the cognitive laboratory approach to apply the fundamental principles derived from cognitive science to our more expansive pretesting procedures. In the current case, however, we found that this conceptual focus was inappropriate for our interviewer review session, and we feel that a strong explicit emphasis on cognition would have been problematic in the pilot field testing exercise (for example, instructing our interviewers to ask spontaneous probes of the type used in laboratory interviews, as described by Willis, et al. (1991), would probably not have been successful). Admittedly, it is possible to provide training, to both interviewers and observers, in the cognitive approach to survey methodology, in order to emphasize the terminology and methods that characterize this approach. Given such training, it would be appropriate to incorporate explicit cognitive features into virtually any form of pretesting. However, interviewers and observers who are not initially accustomed to the cognitive focus do not automatically select this as a frame of reference in performing their evaluations, and forcing this upon them, without adequate training and preparation, may create a number of difficulties.

### References

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

Campanelli, P., Rothgeb, J. M., & Martin, E. (1989). The role of respondent comprehension and interviewer knowledge in CPS labor force classification. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 425-429.

Dippo, C. (1989). The use of cognitive laboratory techniques for investigating memory retrieval errors in retrospective surveys. Proceedings of the International Association of Survey Statisticians, International Statistical Institute, pp. 323-342.

Fowler, F. J. (1989). Coding behavior in pretests to identify unclear questions. Proceedings of the Fifth Conference on Health Survey Research Methods. F. J. Fowler (ed.), pp. 9-12. Washington, D.C.: National Center for Health Services Research and Health Care Technology Assessment.

Gower, A. R., & Dibbs, R. (1989, March). Cognitive research: Designing a respondent-friendly questionnaire for the 1991 Census. Paper presented at the Fifth Annual Research Conference (ARC V), United States Bureau of the Census, Arlington, Virginia.

Jobe, J. B. & Mingay, D. J. (1991). Cognition and survey measurement: History and overview. Applied Cognitive Psychology, 5, 175-192.

Stein, B. J. (1992, May). Subject recruitment and management of a questionnaire design laboratory. Paper presented at the Field Directors Conference, St. Petersburg, Florida.

Willis, G. B. (1991, May). The use of behavior coding to evaluate a draft health-survey questionnaire. Paper presented at the meeting of the American Association for Public Opinion Research, Phoenix, Arizona.

Willis, G. B., Royston, T., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. Applied Cognitive Psychology, 5, 251-267.

The authors would like to thank Dr. Douglas Herrmann, Office of Research and Methodology, National Center for Health Statistics, for his comments on an earlier draft of this paper.