IMPROVING SURVEY QUALITY THROUGH PRETESTING

Theresa J. DeMaio, Jennifer Rothgeb, Jennifer Hess, U.S. Bureau of the Census
Theresa J. DeMaio, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Cognitive, Pretesting

As the vehicle of data collection, the questionnaire is one of the critical components in achieving high quality in a survey. The best of sampling schemes and estimation strategies will not yield accurate data if the answers provided by the respondent are not meaningful.

During the last decade or so, there has been increased emphasis on building quality into the questionnaire design process through pretesting. This has been approached from an operational perspective in Federal government agencies (DeMaio, 1983; DeMaio et al, 1993; Willis, 1994; Dippo and Norwood, 1992) but it has been informed by theoretical work in the areas of cognitive psychology (Tourangeau, 1984; Ericsson and Simon, 1980, 1984) and social psychology (Cannell, Fowler, and Marquis, 1968; Turner and Martin, 1984).

Whereas prior to this time the main contributors to diagnosing questionnaire problems were the questionnaire designers (through their expertise in the subject) and the interviewers (through their experience in administering the questionnaires), the emphasis has shifted in recent years to learning about questionnaire problems from the respondents themselves. This has been predicated on the development of a model of survey response (Tourangeau, 1984; Strack and Martin, 1987) that divides the response process into four stages -- comprehension, retrieval, judgment, and response formulation -- which occur within the respondent and the understanding of which allows researchers to get a grasp on issues that impact the quality of the data collected in surveys.

From a theoretical perspective, research (Sudman, Bradburn, and Schwarz, 1996 ; Schuman and Presser, 1981; Bradburn, Sudman, and Associates,1975) has provided guidance and insight into such questionnaire phenomena as context effects, primacy and recency effects, development of response categories, and social desirability effects based on theories of cognitive and social psychology. This work has general implications in terms of how to structure and sequence the questions and response categories in the initial development of the questionnaire.

However, even informed by the best theoretical research, the operational aspects of questionnaire development and testing should not be neglected. In any particular survey, there may be aspects specific to the population, the subject matter, or the data collection methodology that affect the ability of the questionnaire to perform as intended. The rest of this paper focuses on the operational perspective of improving data quality – the pretesting of an individual questionnaire. The importance of this critical step cannot be underestimated. With all the time, labor, and financial resources that go into the preparation of a survey design and data collection instrument, the last step of pretesting the instrument before it is actually administered may be the deciding factor in whether the survey is successful in meeting its objectives.

In this paper, we describe three of the methods that are used to pretest questionnaires: cognitive interviewing, respondent debriefing, and behavior coding of respondent/interviewer interaction. We recognize that there are several other valuable pretesting methods such as expert panels, questionnaire appraisal coding systems, interviewer debriefings, etc. The methods we discuss were chosen since together they provide information from all three potential sources of measurement error: the questionnaire, interviewer, and respondent. While these three methods are used widely throughout the Federal government and by other data collectors as well, we will concentrate on their use at the Census Bureau. We present brief descriptions and historical summaries of each of these methods, as well as examples of their use, the results they obtain, and how they can be used to improve questionnaires.

Cognitive Interviewing

Cognitive interviews are considered a "laboratory" method because the interviews are conducted one-on-one with a researcher and a subject, and they typically take place in a laboratory, although they can also take place in the respondent's home or in a central location such as a library. Adapted from the method of protocol analysis which was developed to study problem-solving (Ericsson and Simon, 1980; 1984), this method involves having respondents think aloud and verbalize their thought processes as they interpret the survey questions and formulate their answers. In a pure "think aloud" interview, the interviewer is essentially silent during the response process and interacts with the respondent only to issue nondirective probes to say aloud what he/she is thinking or to elaborate on something he/she has said. This pure method is not used universally, however, and in many survey organizations (Willis, 1994; DeMaio et al, 1993), cognitive interviews are a combination of having the respondent think aloud and having the interviewer probe for the respondent's definition of terms or concepts (e.g., "what does the term quit smoking mean to you?") or

interpretation of the question (e.g., "can you tell me in your own words what this question means to you?").

Cognitive interviews can be conducted as either concurrent interviews or retrospective interviews. In concurrent interviews, respondents describe their thoughts while answering the questions. In retrospective interviews, the respondent first completes the interview, similar to the conditions under which most survey respondents would complete the interview task. Following the interview, the respondent and interviewer review the survey responses and the respondent is asked about the process used to generate his/her answers. Interviewers can use general probes (e.g., "tell me what you were thinking when you ...") or very specific probes (e.g., "Why did you report the $142 payment as child support?") to guide the think aloud process. The concurrent think aloud has the advantage of capturing the information at the time it is first available to the respondent; however, it may bias responses to questions later in the interview. Retrospective techniques provide an unbiased means for capturing the data, while still preserving the opportunity to focus on general or specific questions concerning the interview. However, the respondent may not be able to recall his/her thought processes when asked about them at the end of the interview rather than after each question. Concurrent and retrospective techniques may capture different kinds of data, especially in the context of a self-administered interview where issues of motivation take on more importance than in an interviewer-administered interview (Redline et al, 1998).

The cognitive interview method is essentially a qualitative research tool. The number of interviews is small (typically 20 or less). The respondents are not scientifically selected; rather, subjects are purposively recruited to permit testing of separate parts of the questionnaire (for example, sections for smokers, former smokers, and non-smokers [DeMaio et al, 1991]) or the same part of the questionnaire by different types of people (for example, questions about race by persons of different races [Gerber, de la Puente, and Levin, undated]). Although the results cannot be generalized to a larger universe, the method is very useful for providing input about how respondents actually formulate their answers and the kinds of errors they introduce that is not available through any other method.

Twenty cognitive interviews using the concurrent think aloud method with probes were used to test new questions on disability proposed for the Census 2000 Dress Rehearsal. Because the census is administered to the entire population, the performance of both disabled persons and non-disabled persons is important to the quality of data collected by these questions. Using a targeted recruitment strategy for cognitive interviews, respondents who were disabled, respondents who were reporting for disabled household members, and non-disabled respondents were recruited to evaluate their understandings of the questions and their question-answering strategies (DeMaio and Wellens, 1997). This allowed a much more highly concentrated test of these questions among the disabled population than would be possible by a random selection method.

An interesting feature of this research (DeMaio and Wellens, 1997) was that two different question versions were tested, thus allowing a comparative evaluation of question series. One version of this two-question self-administered series (see Attachment 1A) suffered primarily from problems of format: the response categories for reporting "no disability" were in the left-hand column for the first question and the right-hand column for the second question. Respondents who thought they had learned the pattern of responses in the first question incorrectly reported household members as disabled in the second question.

The second question series (see Attachment 1B) revealed more serious problems of interpretation. One question which was designed to reveal sensory disabilities ("Because of a physical, mental, or emotional limitation lasting six months or more, does this person have any difficulty in doing any of the activities listed below? "Talk, see (with glasses), or hear" revealed errors in reporting in the vision and speech areas. Respondents were confused about whether the "see (with glasses)" phrase meant "difficulty seeing with their glasses on" or "difficulty seeing and needed glasses." Many respondents who wore glasses incorrectly reported that they had problems with this activity because they wore glasses. This is an example of measurement error in the direction of overreporting disability. In contrast, the respondents underreported difficulties with speech. Two respondents – one with epilepsy and one with a brain injury from a fall – did not report difficulties with their speech that seemed quite apparent to the interviewers. Also, a parent reported that her child had no difficulty talking in response to the question, but then later noted that his slurred speech requires him to go to a speech therapist. She thought about the question as an either/or proposition–either her son could speak or he couldn't. But she did not think the question was referring to the quality of his speech.

This level of detail about the interpretation of the questions, response categories, and format of the questionnaire is invaluable for providing information about ways in which the questions are not achieving the goals of the questionnaire designers and are not yielding high quality data.

Respondent Debriefing

In contrast to cognitive interviewing, which is conducted in a laboratory setting, respondent debriefing is incorporated into the actual data collection method of the survey. It can be included as part of a survey pretest, to provide input for revision for the production survey, or it can be included in the actual survey to provide input for the next administration of a continuing survey. Respondent debriefing involves incorporating follow-up questions in a field test interview to gain a better understanding of how respondents interpret questions asked of them. The technique was originally used many years ago (Belson, 1981), but has gained in popularity in recent years (Fowler and Roman, 1992; Oksenberg et al, 1991; Esposito et al, 1991; Esposito et al, 1992; Esposito and Rothgeb, 1997; Nelson, 1985; Hess and Singer, 1995). This technique is sometimes referred to in the literature as special probes (Oksenberg et al, 1991) or frame of reference probing (DeMaio, 1983).

The primary objective of respondent debriefings is to determine whether concepts and questions are understood by respondents in the same way that the survey designers intend. In addition, respondent debriefings can be quite useful in determining the reason for respondent misunderstandings. Sometimes, results of respondent debriefing show that a question is unnecessary and does not need to be included in the final questionnaire. Conversely, it may be discovered that additional questions need to be included in the final questionnaire in order to better operationalize the concept of interest. Finally, the data may show that concepts or questions cause confusion or misunderstanding as far as the intended meaning is concerned. In any of these cases, changes can be incorporated into the final questionnaire to reduce measurement error. Respondent debriefing can also be used to obtain information on the respondent's perception of task difficulty or question sensitivity. Items determined to be difficult or sensitive can be modified (or eliminated) to reduce difficulty and/or sensitivity.

Several different methods of collecting respondent debriefing information exist. Follow-up probes are administered, generally at the end of the interview so as not to interfere with the context of the interview itself, to elicit information about how respondents are interpreting the question and, as a result, how they are arriving at their answers. (This can be done with interviewer-administered questionnaires as well as self-administered questionnaires.) Probes can be either open-ended or closed-ended. In the former, respondents can be asked how they developed their response to the target question (e.g., what did they include or exclude when formulating their response). For example this method was used to determine whether respondents included other kinds of reading material when asked whether they read any novels within the past 12 months (Fowler and Roman, 1992). With closed-ended probes, structured questions can include specific response alternatives that respondents might have employed, and ask which one was the one they used. For example, respondents can be offered various reference periods and asked which they used in formulating their response to determine if the intended reference period is the one being used (Hess and Singer, 1995).

Follow-up probes generally refer to the respondent's particular situation, and how he/she interprets and answers a question that applies to him/her. This can be limiting, in that the size of the field test may limit the collection of data that apply to relatively rare types of situations. One way to collect information about rare situations, or a wide variety of situations rather than simply the respondent's own, is through the vignette. Vignettes present hypothetical situations and ask for the respondents' classification of the situation based on their interpretation of the concept. For example, the decennial census uses a complicated and sometimes counterintuitive set of rules to establish whether persons should be considered household members for purposes of the census count. Vignettes have been used to elicit respondents' interpretations of whether persons in hypothetical scenarios should be included when the hypothetical household is rostered in the census (Gerber, 1994; Gerber, Wellens, and Keeley, 1996).

All three of these types of respondent debriefings were used during the redesign of the Current Population Survey (CPS) in the early 1990s. The CPS, which is the large-scale monthly Federal survey from which the national and state-level unemployment rates are derived, went through an extensive process to revise the questionnaire. The goal was to update the questionnaire to improve reporting in areas where poor reporting was suspected to exist (Bregger and Dippo, 1993). As part of the research program, three large field tests were conducted using alternative questionnaires, and follow-up probes and vignettes were incorporated to understand how respondents were interpreting concepts of critical importance to the survey, such as "work," "main job," and "business."[1]

According to Esposito et al (1993), follow-up questions were used for five reasons: "(1) to establish whether there were any misunderstandings of terms or phrases used in the main survey; (2) to ascertain the extent to which respondents' understandings of questions and concepts were consistent with official definitions; (3) to evaluate whether some questions in the main survey

[1]The redesign of the CPS also included a number of other pretesting methods as evaluation tools. See Esposito et al (1993) for a fuller discussion.

were superfluous; (4) to examine whether alternate versions of a question did a better job of identifying or measuring specific activities; and (5) to construct comparable subsets of respondents from different questionnaire versions to allow comparative analyses.(pp.18-19)"

Open-ended questions were used to establish, for persons who had more than one job, how the main job was decided upon (e.g., "You mentioned earlier that you had more than one job. How did you decide which job was your MAIN job?"). The interviewers field-coded the responses, and the results showed that only 63 percent of the respondents reported as their main job the one at which they worked the most hours, which is the intent of the survey designers. Thus, this was a candidate for revision by including the definition of the "main job" concept in the CPS questionnaire.

The closed-ended approach was used to evaluate whether new question wording was an improvement over the old version in capturing reports of casual employment, that is, informal and/or irregular work arrangements. Follow-up probes were coordinated with the question asking whether the person had worked last week, which was asked in three different ways in different questionnaire panels. Persons who reported not working last week (in any of the panels) were asked if they had done any casual work during the past week and asked to describe the previously unreported work activity. The results showed that each of the questionnaire versions missed some casual employment, but the amounts were small (in the 1-2 percent range) and that one of the versions did marginally better at capturing such employment (Esposito and Rothgeb, 1997). This suggested that other factors in evaluating the "work last week" question should be given priority in deciding which one to use.

Vignettes were used to learn about respondents' interpretation of the concept of "work." A series of hypothetical situations that focused on ambiguous ones such as volunteer work, work at home for a family business, preparing to start a business, and casual labor were presented, and respondents were asked whether the person in the vignette should be reported as working. Evaluation of the vignette reports (Martin, Campanelli, and Fay, 1991; Martin and Polivka, 1995) was used to determine how broadly or narrowly respondents viewed the concept, and which aspects of the definition were most poorly understood. Comparison of the performance of the vignettes across respondents who were administered different versions of the "work last week" question provided information about problems of comprehension and question wording. In addition, comparison of responses to the same vignettes and the same wording of the "work last week" question by different respondents at different points in time revealed similar results, suggesting that this is a robust tool for measuring the meaning of key survey constructs (Martin and Polivka, 1995).

Data obtained from respondent debriefings (and other questionnaire evaluation methods) conducted during the testing of the CPS questionnaire demonstrate that "the revised questions are more clearly understood by respondents and the potential for labor force misclassification is reduced" (Rothgeb et al, 1994).

Behavior Coding and Analysis

Another useful method for evaluating the quality of pretest data for interviewer-administered surveys is behavior coding, or coding of the interchange between the interviewer and respondent. The coding is done in a systematic way on a case-by-case basis to capture specific aspects of how the interviewer asked the question and how the respondent answered. This method was first used in surveys to monitor and evaluate interviewer performance (Cannell et al, 1975) and subsequently as a tool to evaluate the question-answer process more generally (Mathiowetz and Cannell, 1980; Morton-Williams and Sykes, 1984; Oksenberg et al, 1991) and to assess the effects of interviewer behavior on response variance (Groves et al, 1980). The method is flexible in that the coding scheme used in any particular application can be adapted to meet specific priorities of the questionnaire designers. A narrow or broad range of interviewer and respondent behaviors can be captured, depending on how much time is available and how much detail is required. One or more rounds of interaction between the respondent and interviewer can be coded and analyzed as well.

The guiding principle behind the use of behavior coding for questionnaire evaluation is that the behaviors of the interviewer and respondent provide insight into problems with the question wording or questionnaire format. For example, if the interviewer does not read the question exactly as it is worded, there could be a problem with awkward wording. If the interviewer omits a question entirely, there may be a problem with the skip instructions or the way a paper questionnaire is formatted. If the respondent interrupts before the interviewer has finished reading the question, perhaps the question is too long. Or if the respondent asks for clarification, there may be a problem with the definition of a term, concept, or the intent of the question.

While behavior coding pinpoints the location of questionnaire problems, it does not necessarily identify the cause of the problems. The behavior coders, however, can be used as a source of information about why problems occurred. The coders can be instructed to provide written comments in problem situations, and they

can be debriefed after all the interviews have been coded. Their recent experience in listening to and coding the interviews may suggest particular features of the questions that are problematic.

Behavior coding was one of the methods used to develop and evaluate the Food Security Supplement to the CPS, administered in April 1995.[2] A field pretest was conducted in August 1994. Questions were revised and an evaluation of the April 1995 Food Security Supplement to the CPS, also called a quality assessment, was conducted to assess how well the revised questions were working. Behavior coding was used as an evaluation methodology during both the pretest and the quality assessment.

Examination of behavior coding rates was used to identify the most problematic questions and suggest revisions to the items. In the food expenditures section of the pretest questionnaire, the following questions were very difficult for respondents:

1. "How much did your household spend for food at a supermarket last week, NOT counting money spent on nonfood items, such as cleaning or paper products, pet food or cigarettes?"
2. "How much did your household spend for food at convenience stores or grocery stores other than a supermarket last week?"

Only 70 percent of respondents provided adequate or qualified answers to the first item above, and written notes provided by coders indicated several types of problems. Some respondents focused on the exclusionary statement at the end of the question ("...NOT counting money spent on nonfood items, such as cleaning or paper products, pet food, or cigarettes?"), rather than on the question itself and reported, for example, that they don't smoke. There were also several requests for clarification regarding the reference period and what to include and exclude from the calculation indicating, perhaps, that the question was too long. The written notes from behavior coding also indicated that the terms "supermarket" and "grocery store" were synonymous for some respondents; thus, the questions seemed redundant and confusing to them.

We revised the question wording for the production survey by combining the questions about purchases at supermarkets and grocery stores in one item and separating estimates of expenditures on nonfood items into a separate item, as shown below:

1. "How much did your household spend at supermarkets and grocery stores last week?"
2. "How much of the (fill with dollar amount from item 1) was for nonfood items, such as cleaning or paper products?"

Behavior coding in the production survey indicated that the revised items still caused problems for respondents. Only 76 percent (N=121) gave adequate or qualified answers. The written notes indicated that some respondents interpreted the question as asking for an average or usual amount, rather than a specific amount spent last week. Respondents also included monies for food purchases at other places besides grocery stores and supermarkets. This was the first in a series of questions designed to elicit reports of actual (last week) and usual (in general) reports of money spent for food at all types of places. Other items in the series were even more problematic, with some showing that less than half the respondents gave adequate or qualified answers. The quality assessment indicated that the food expenditure series of questions was still causing problems for respondents. Several recommendations for revisions were made (see Hess, Singer, and Ciochetto, 1995).

In addition to using behavior coding to evaluate the questions, Hess and Singer (1996) attempted to compare behavior coding results with results of an independent reinterview conducted to evaluate the consistency of response on the same questions to two independent measurements. This is an innovative idea for documenting the diagnostic utility of behavior coding as a pretest method. Behavior coding results are compared question by question with the results of a reinterview in which the same questions were asked approximately one week after the original supplement was administered.[3] In the Food Security Supplement, the results showed that while interviewer behaviors were not significantly associated with reinterview problems, respondent behaviors were. (Recall that the quality of interviewer question-reading in this survey was very high.) Thus, questions that showed a high level of respondent problems in the behavior coding also had a lower level of data quality as evidenced by the Index of Inconsistency. While this work is suggestive rather than definitive because of the imprecise level of correspondence between the behavior coding respondents and the reinterview

[2]See Singer and Hess (1994) and Hess, Singer, and Ciochetto (1996) for a description of the questionnaire evaluation process for the pretest and the production survey, respectively.

[3]The comparison is between aggregate results of the reinterview and the behavior coding. Responses could not be matched for the same respondents due to the design of the reinterview. Quality of the reinterview response is measured by the Index of Inconsistency, which indicates that a question has a low, moderate or high level of response variance (U.S. Bureau of the Census, 1993).

respondents, it nevertheless provides evidence of an association between these two methods. It also provides the direction for further research in this area, in terms of characteristics of questions for which behavior coding is more and less useful as a diagnostic measure.

Discussion

This paper shows that there are a wide variety of pretesting tools that can be used to improve the quality of survey data. These methods are both qualitative, as in the case of cognitive interviews, and quantitative, as in the case of respondent debriefing and behavior coding. They are typically used at different points in the survey development process, with cognitive interviews coming earlier and being used to refine the questionnaire before administering it in a larger-scale field test that incorporates respondent debriefing and behavior coding. These are not hard and fast rules, since vignettes have also been used in conjunction with cognitive interviews and quantitative analysis has also been conducted on cognitive interview data.

Although we have presented the results of respondent debriefing and behavior coding as individual case studies, these methods are in fact complementary. Behavior coding of interviewer-respondent interactions is useful to demonstrate that a problem exists with survey questions. On the other hand, respondent debriefing can be used to provide information about the reasons for the problematic item(s). In combination, these two methods, used in the conduct of either pretests or actual field administration, provide valuable information about whether, how often, and (in the best of circumstances) how respondents misunderstand survey questions and provide a basis on which the questions can be revised, either for a single data collection period or the next administration of a continuing data collection.

In this paper we have focused on methods by which to assess survey quality during questionnaire pretesting. It is encouraging that increasingly more research organizations are realizing the benefits of improved data quality when investing the resources in such pretesting. What seems to be ignored, however, is assessing data quality once the questionnaire is administered in the production survey. Frequently, question revisions resulting from question pretesting do not get retested prior to inclusion in the production survey. As a result, it is rarely known whether the expected benefit from the revision is realized. (We have presented one instance, the Food Security Supplement, in which the pretest revisions were evaluated in the production survey. However, this experience is rare.)

In recent years, researchers have described how question evaluation methods (such as those described in this paper) used during pretesting can also be used as quality assessment techniques during production surveys, as well as other quality assessment measures such as reinterview (Esposito and Rothgeb, 1997). In order to adequately (and accurately) measure survey quality (from the questionnaire design perspective), it is necessary that quality assessment measures be collected during the production survey. We encourage survey organizations, particularly those with large-scale recurring surveys, to implement quality assessment programs.

"This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion."

References

Belson, W. *The Design and Understanding of Survey Questions*. Aldershot, England: Gower, 1981.

Bradburn, N., Sudman, S., and Associates. *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass Publishers, 1979.

Bregger, J.E., and Dippo, C.S. "Overhauling the Current Population Survey: Why is it Necessary to Change?" *Monthly Labor Review*, 1993,Vol. 116, No. 9, pp. 3-9.

Cannell, C.F., Fowler, F.J., and Marquis, K. *The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting in Household Interviews*. Vital Health and Statistics, Series 2, Number 26, Washington, DC, Government Printing Office, 1968.

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

DeMaio, T. (ed). *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10, Washington, DC: Office of Management and Budget, 1983.

DeMaio, T., Mathiowetz, N., Rothgeb, J., Beach, M.E., and Durant, S. *Protocol for Pretesting Demographic Surveys at the Census Bureau*, Census Bureau Monograph, Washington, DC: U.S. Bureau of the Census, 1993.

DeMaio, T. and Wellens, T. "Cognitive Evaluation of Proposed Disability Questions for the 1998 Dress Rehearsal," unpublished Census Bureau report, 1997.

DeMaio , T., Ciochetto, S., Sewell, L., Beach, M.E., and Glover, T. "Report on Results of Cognitive Interviewing for the CPS Tobacco Use Supplement for the ASSIST Evaluation," Unpublished Census Bureau report, 1991.

Dippo, C.S. and Norwood, J.L. "A Review of Research at the Bureau of Labor Statistics" in J.M. Tanur (ed), *Questions about Questions*, New York: Russell Sage Foundation, 1992.

Ericsson, K.A. and Simon, H.A. *Protocol Analysis*. Massachusetts: The MIT Press, 1984.

Ericsson, K.A. and Simon, H.A. "Verbal Reports as Data," *Psychological Review*, 87:215-251, 1980.

Esposito, J.L., Campanelli, P.C., Rothgeb, J., and Polivka, A.E. "Determining Which Questions Are Best: Methodologies for Evaluating Survey Questions," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1991, pp. 46-55.

Esposito, J.L., Rothgeb, J.M., Polivka, A.E. Hess, J., and Campanelli, C. "Methodologies for Evaluating Survey Questions: Some Lessons from the Redesign of the Current Population Survey," paper presented at the International Conference on Social Science Methodology, Trento, Italy, 1993.

Esposito, J.L,. Rothgeb, J.M., and Campanelli, P.C. "The Utility and Flexibility of Behavior Coding as a Methodology For Evaluating Questionnaires," paper presented at the American Association for Public Opinion Research, 1994.

Esposito, J.L. and Rothgeb, J.M. "Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment." in L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), *Survey Measurement and Process Quality*. New York: John Wiley and Sons, Inc. 1997.

Fowler, F. and Roman, T. "A Study of the Approaches to Survey Question Evaluation," Center for Survey Research, University of Massachusetts, 1992.

Gerber, E. "Hidden Assumptions: The Use of Vignettes in Cognitive Interviewing," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1996, pp. 1269-1274.

Gerber, E., de la Puente, M., and Levin, M. "Race, Identity, and New Question Options: Final Report on Cognitive Research on Race and Ethnicity," Unpublished Census Bureau report, undated.

Gerber, E., Wellens, T., and Keeley, C. "Who Lives Here?: The Use of Vignettes in Household Roster Research," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1996, pp. 962-967.

Groves, R., Berry, M, and Mathiowetz, N. "The Process of Interviewer Variability: Evidence from Telephone Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1980, pp. 519-524.

Hess, J. and Singer, E. "The Role of Respondent Debriefing Questions in Questionnaire Development," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1995, pp. 1075-1080.

Hess, J. and Singer, E. "Predicting Test-Retest Reliability from Behavior Coding," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1996. pp. 1004-1009.

Hess, J., Singer, E., and Ciochetto, S. "Evaluation of the April 1995 Food Security Supplement to the Current Population Survey, internal Census Bureau report, 1996.

Martin, E.A., Campanelli, P.C., and Fay, R.E. "An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of 'Work' in the Current Population Survey," *Statistician*, 40: 265-276, 1991.

Martin, E. and Polivka, A.E. "Diagnostics for Redesigning Questionnaires," *Public Opinion Quarterly*, 59: 546-567, 1995.

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

Nelson, D. "Informal Testing as Means of Questionnaire Development," *Journal of Official Statistics*, 1: 179-188, 1985.

Oksenberg, L., Cannell, C., and Kalton, G. "New Strategies for Pretesting Survey Questions," *Journal of Official Statistics*, 7:349-365, 1991.

Redline, C., Smiley, R., Lee, M., DeMaio, T., and Dillman, D. "Beyond Concurrent Interviews: An Evaluation of Cognitive Interviewing Techniques for

Self-Administered Questionnaires," paper presented at the Annual Meetings of the American Association for Public Opinion Research, 1998.

Rothgeb, J. "Summary Report -- MDS Phase II." Unpublished Census Bureau memorandum, 1982.

Rothgeb, J., Cohany, S., Esposito, J., Hess, J., Polivka, A., and Shoemaker, H. "Revisions to the CPS Questionnaire: Effects on Data Quality." Report submitted to the Bureau of Labor Statistics, April 6, 1994.

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

Strack, F, and Martin, L.L., "Thinking, Judging and Communicating: A Process Account of Context Effects in Attitude Sulrveys: in H.-J. Hippler, N. Schwarz, and S. Sudman (eds.), *Social Information Processing and Survey Methodology*. New York: Springer-Verlag, 1987.

Sudman, S., Bradburn, N.M., and Schwarz, N. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*, San Francisco: Jossey-Bass Publishers, 1996.

Tourangeau, R. "Cognitive Science and Survey Methods," in T. Jabine et al (eds.) *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC: National Academy Press, 1984.

Turner, C. and Martin, E. (eds.) *Surveying Subjective Phenomena*, Vols. 1 and 2, New York: Russell Sage, 1984.

U.S. Bureau of the Census. *Content Reinterview Survey: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview*, 1990 CHP-E-1. 1993.

Willis, G. *Cognitive Interviewing and Questionnaire Design: A Training Manual*, Cognitive Methods Staff Working Paper Series, No. 7, Hyattsville, MD: U.S. National Center for Health Statistics, Office of Research and Methodology, 1994.

ATTACHMENT 1A

VERSION A

Mark the category that best describes this person's usual ability to perform the following activities:

	No difficulty	Some difficulty	Great difficulty or unable
a. Perform mental tasks such as learning, remembering, concentrating	☐	☐	☐
b. Dress, bathe, and get around inside the home without help from another person	☐	☐	☐
c. *Answer if person is 16 YEARS OLD OR OVER* – Go outside the home alone to shop or visit a doctor's office	☐	☐	☐

Does this person have any of the following long-lasting conditions —

	Yes	No
a. Blindness or a severe vision impairment?	☐	☐
b. Deafness or a severe hearing impairment?	☐	☐
c. A condition that substantially limits one or more basic physical activities such as walking, climbing stairs, reaching, lifting, or carrying?	☐	☐

ATTACHMENT 1B

VERSION B

Because of a physical, mental, or emotional limitation lasting 6 months or more, does this person have any difficulty in doing any of the activities listed below?

	Yes	No
a. Learn, remember, or concentrate	☐	☐
b. Talk, see (with glasses), or hear	☐	☐
c. Walk 3 blocks or lift a bag of groceries . .	☐	☐
Answer if person is 16 YEARS OLD OR OVER —		
d. Work or keep house	☐	☐
e. Go outside the home alone to shop or visit a doctor's office	☐	☐
f. Dress, bathe, or get around inside the home	☐	☐