DETERMINING WHICH QUESTIONS ARE BEST: METHODOLOGIES FOR EVALUATING SURVEY QUESTIONS

James L. Esposito, Bureau of Labor Statistics
Pamela C. Campanelli and Jennifer Rothgeb, Bureau of the Census
Anne E. Polivka, Bureau of Labor Statistics
James L. Esposito, GAO Building, 441 G Street, NW, Room 2127, Washington, DC 20212

Key Words: CPS pretest, survey methodology, CATI

Introduction

While survey researchers agree on the need for pretesting questionnaires, there is no accepted methodological framework that outlines how one might accomplish such a task in a comprehensive and efficient manner. Recently, however, there has been a concerted effort by survey researchers to document ways of evaluating survey questions and to systematize current pretest methodologies (e.g., Belson, 1981; Cannell et al., 1989; Converse and Presser, 1986; DeMaio, 1983; Oksenberg, Cannell, and Kalton, 1991; Nelson, 1985). For example, in a report entitled, *Approaches to Developing Questionnaires*, DeMaio (1983) describes various methods and techniques for evaluating survey questions. These include, among others, interview monitoring, frame-of-reference probing, and interviewer debriefing. These three general methodologies were adapted for use in the current pretest and are discussed very briefly below.

The monitoring and coding of exchanges between interviewers and respondents during the course of an actual interview is a method that has been used by survey researchers to evaluate interviewer performance (Cannell, Lawson, and Hausser, 1975) and to investigate the question-and-answer process more generally (Dijkstra et al., 1985; Morton-Williams, 1979; Marquis, 1969). One specific technique, *behavioral coding*, involves monitoring or tape recording actual interviews and keeping a quantitative record of those questions which interviewers fail to read correctly and those for which respondents have problems. Cannell et al. (1989) suggest that behavior coding is useful for identifying problematic questions and for providing data on both the prevalence and nature of such problems.

Frame-of-reference probing refers to using a series of additional probes to ascertain whether certain words, phrases, or situations are understood by respondents in the manner intended by the questionnaire designer. This technique has been used extensively by Belson (1981) who developed a "question testing" method to look at potential respondent comprehension failures (see also Cantril and Fried, 1944; Ferber, 1956; Oksenberg, Cannell, and Kalton, 1991; Nelson, 1985; Schuman, 1966.) Various techniques (e.g., paraphrasing, think-aloud methods) arising from the movement to develop cognitive aspects of survey methodology (Jobe and Mingay, 1991; Tanur and Fienberg, 1990) are also designed to probe respondents' frame-of-reference in answering survey questions. Another technique which could be classified under the general heading of frame-of-reference probing is the use of field-based respondent debriefing studies. This technique involves moving the cognitive laboratory into the "field" so that information can be obtained from a larger, more representative sample of respondents (e.g., Campanelli et al, 1989; Martin et al. 1986).

Interviewer debriefings (e.g., focus groups) are useful when pretesting questionnaires because interviewers are in a unique position to evaluate the merits of survey questions (e.g., see Converse and Schuman, 1974). Not only do they obtain useful feedback from respondents in the course of administering questionnaires, but more experienced interviewers can draw on their accumulated knowledge of survey interactions to identify in advance which items are likely to be problematic.

Although evidence suggests that each of these techniques can lead to improvements in survey measurement, their relative usefulness is unclear. In this paper, we focus on what we have learned about the strengths and weaknesses of these techniques regarding ease of implementation, interpretation, and effectiveness at identifying problematic questions.

Background

As part of the effort to redesign the Current Population Survey (CPS) questionnaire, the Census Bureau--in collaboration with the Bureau of Labor Statistics--is conducting a Computer-Assisted Telephone Interview (CATI) Random Digit Dialing (RDD) test at its interviewing facility in Hagerstown, MD. The first phase of this CATI/RDD test (July 1990 to January 1991) involved approximately 72,000 persons. The second phase (July to October 1991) involved approximately 30,000 persons. The purpose of phase one was to compare the current version of the CPS questionnaire ("A") with two new alternative versions ("B" and "C"), which were developed on the basis of earlier laboratory and field research (e.g., BLS, 1988; Campanelli et al., 1989; Fracasso, 1989; Palmisano, 1989). The principal product of the first phase was a single alternative questionnaire ("D"), which comprised the *best* questions from versions A, B, and C, as well as any new questions deemed necessary due to the results of the first phase. In the second phase of the pretest, the current CPS questionnaire was tested against version D with the goal of producing a fully redesigned CPS questionnaire for the 1990's.

The two alternative questionnaires (B and C) differed from the current CPS questionnaire (A) in that questions were revised, or new questions added, in an effort to: (1) better operationalize existing definitions, (2) implement preestablished definitional changes within the labor force classifications (e.g., discouraged workers), (3) expand or elaborate on the type of labor force data being collected (e.g., number of multiple job holders), and (4) better utilize the capabilities of computer-assisted interviewing in an effort to improve data quality and, in some cases, reduce cost and burden. [For further information on the CPS questionnaire redesign project, see Copeland and Rothgeb (1990) and Rothgeb et al. (1991).]

Methods

To facilitate the decision-making process for selecting items to appear in version D, several different methods and techniques were used: (1) systematically coded interviewer-respondent interactions, (2) interviewer debriefings using a standardized questionnaire and focus groups, (3) field-based respondent debriefings using answer-keyed follow-up questions and vignettes, and (4) nonresponse and response-distribution analyses.

1. Systematically coded interviewer-respondent interactions. Using a specially developed form inspired by the work of Cannell et al. (1989), we coded interviewer-respondent interactions at the Census Bureau's CATI facility over a period of six months (July through December, 1990). Our procedures were designed to allow the coding of interviewer-respondent exchanges *during an actual interview*. In this respect, our use of the behavior-coding technique differs substantially from procedures used by others (e.g., Cannell et al., 1989; Oksenberg, Cannell, and Kalton, 1991; Shepard and Vincent, 1991).

Our coding form allowed for the recording of repeated "levels of exchange" for any question. A coded interviewer behavior followed by a coded respondent behavior constituted a single, complete *level of exchange*. Monitors noted whether interviewers read a question exactly as worded, with a slight change in wording, or with a major change in wording. Our working assumption was that a major change in wording would substantially alter the meaning of the question. Even minor deviations in question wording (e.g., leaving out key words, such as "last week") were coded as major changes if they altered the meaning or intent of the question. Any first-level exchange where the interviewer verified relevant information with the respondent--in lieu of reading the survey question--was coded as a *verify* in the interviewer-behavior column labeled "probe, feedback, or verify". For the respondent, we distinguished among the following behaviors: Gives adequate answer, gives qualified answer, gives inadequate answer, asks for clarification, interrupts, doesn't know, or refuses to answer.

Six researchers from BLS and Census served as monitors. *Dual coding*--that is, two monitors coding the same interview independently--was conducted twice; once in July, as a training exercise, and again in October, to obtain a measure of "inter-monitor" reliability. In general, our inter-monitor reliability was moderately high. Over all questions, there was 81% agreement on the full range of interviewer behavior codes and 76% agreement on the full range of respondent behavior codes.

A total of 229 household interviews were monitored, yielding data on 483 persons. Over all household members and questions, 4646 first-level exchanges were observed. Though a fairly large number, these 4646 interactions were spread over 201 questions on three questionnaire versions. The main method of examining these data consisted of looking at the first-level exchanges and comparing the percentage of times a behavior occurred for particular questions that were comparable across versions A, B, and C. The base was the number of times the question was asked. For statistical purposes, a threshold was established that a question had to be asked at least 20 times in order to be included in our analyses. This limited our comparisons to 57 questions. These questions were then grouped into 22 across-version *comparison sets*; that is, questions measuring a given topic on one version of the questionnaire were grouped with their counterparts from the other versions--where these were available. For these analyses, we examined tables containing all of the behavior codes for interviewers (e.g., percent exact reading, percent slight change, etc.) and for respondents (e.g., percent giving adequate answer, percent giving qualified answer, etc.). Of the 44 possible comparisons involving interviewer behaviors and respondent behaviors, only six were statistically significant: two of these occurred in places where question wording did not differ by version (and thus were anomalous), two occurred in places where the universe of people who received the question on each version were different enough to prevent a valid comparison, and one occurred in a place where the "comparable" questions were of different question types. This left only one meaningfully significant result.

Overall, interviewers read questions exactly as worded, or correctly verified a prior response, 95% of the time for version A questions (n=1120), 96% of the time for B questions (n=1452), and 93% of the time for C questions (n=1986). These rates of interviewer performance, while higher than results from monitoring studies of in-person interviews (e.g., Bradburn et al., 1979; Brenner, 1982; Marquis, 1971) are consistent with findings from centralized telephone facility studies (e.g., Mathiowetz and Cannell, 1980; Oksenberg, 1981; Presser and Zhao, 1990). On their initial responses, respondents gave an adequate answer 82% of the time for version A questions (n=1073), 85% of the time for B questions (n=1380), and 85% of the time for C questions (n=1854).

Although the variant of behavior coding used for this pretest proved to be a laborious and time-intensive methodology (e.g., coding interactions, creating data files) relative to other methods, it was useful in identifying problematic questions and series. Given the specific objective of the CATI/RDD test, however (i.e., to select the best questions from versions A, B, and C for inclusion in version D), these data were not as useful as we might have hoped. First, some of the series identified as problematic (e.g., industry and occupation) were ones we already knew were problematic from prior research. Second, for some question sets, sample sizes were too small to detect significant differences. It is also worth noting that, in some cases, different versions of a particular question did not vary substantially in content--thus contributing to the problem of detecting differences among a set of comparable questions. Finally, the behavior-coding methodology does not help us to distinguish between adequate answers that are valid and those that are not valid. An adequate answer is one that an interviewer **can code**, given a set of response categories, or one that meets the objective of the question from the interviewer's perspective (as in the case of open-ended questions). However, there is no guarantee that a coded/recorded answer is a valid answer. If the validity of these questions can be assessed at all, it can only be assessed using other methods (e.g., respondent debriefing, response distribution analyses, record checks).

The coding of interviewer-respondent exchanges represents a relatively objective method of identifying questionnnaire items that are causing problems for interviewers and respondents, but it does not inform the survey researcher as to what the reasons for these problems might be. For this information, the researcher must turn to the survey participants themselves. In the next two sections, we describe and discuss the methods we used for debriefing interviewers and respondents.

2. Interviewer debriefing. Our research utilized two aspects of interviewer debriefing: (1) completion of a self-administered debriefing questionnaire, and (2) participation in a focus group discussion with other interviewers. The self-administered questionnaire was distributed to all CATI/RDD interviewers in September of 1990, about 10 weeks after the start of phase one. Eighty-eight percent of the interviewers (68 of 77) returned a completed questionnaire. Six focus group sessions, with 8-10 interviewers each, were then conducted over a three-month period (two each in September, October, and November, 1990).

Although the two aspects of interviewer debriefing utilized different formats, they sought to collect similar information and, as a result, shared a similar underlying structure. Both the questionnaire and the moderator's focus-group guidelines were structured to proceed from interviewers' general preferences for a particular questionnaire version to their specific evaluations of a particular question or series of questions. For example, interviewers were asked: (1) which questionnaire version flowed the best/worst, (2) which series of questions, and later which single question, they thought would be most difficult to answer as a respondent, and (3) which single question respondents refused to answer most often. In addition to the above, interviewers were asked which concepts or terms they felt were most commonly misunderstood or misinterpreted by respondents.

Aside from ease of administration, the principal strength of the interviewer debriefing methods we used was that they enabled us to gain valuable information from interviewers regarding such things as preferences for alternate questionnaires, identification of troublesome questions/series (e.g., earnings) and difficult-to-understand concepts (e.g., "compensation", "private company"), and reactions to new data collection techniques (e.g., dependent interviewing). For example, some interviewers did not like the following question: "Last week, did ... do any work at all? Include work for pay or other types of compensation." They pointed out that this question tends to be problematic for some respondents because of the uncertainty of what is meant by the phrase "other types of compensation" and because the word "compensation" strikes some as a bit too intellectual. Another set of questions that were identified as problematic appear in the industry and occupation (I/O) series:

A. "For whom did ... work?"
B. "What is the name of ...'s employer?"
C. "What is the name of the company for which you work?"

Although these questions were clear and simple enough for interviewers to ask, they seemed to generate resistance from some respondents who found them intrusive.

The format used in conducting focus groups allowed interviewers to discuss important problem areas in depth and to brainstorm possible solutions. For example, in the debriefing questionnaires, interviewers clearly expressed a preference for the use of *dependent interviewing* for industry and occupation data (i.e., when the respondent's previous month's answers are displayed on the CATI screen for the purpose of verifying or updating the information this month). In the focus groups, however, interviewers pointed out some of the shortcomings of this collection strategy. For example, in their attempts to explain why they have to ask the same labor force questions every month, some interviewers tell annoyed respondents that their previous data are not available for this month's interview due to confidentiality safeguards. Though true for most of the data being collected on the CPS, this explanation seems contradictory to respondents when the prior month's data are automatically retrieved for dependent I/O questions. This problem can be easily resolved during training by providing interviewers with guidance on how to explain dependent-interviewing procedures to respondents.

In addition to its strengths, there are a number of weaknesses associated with the use of interviewer debriefing methods. First, insofar as these methods required interviewers to interpret the thoughts and/or behavior of respondents, such data may have been colored by interviewers' own experience and expectations. For example, some interviewers may have reported that respondents have trouble interpreting the phrase "other types of compensation" because of their own difficulties with the phrase, or because they recall a particularly salient interview when they were lambasted by a respondent about the use of such highbrow terminology. In such cases, interviewers may not be making judgments on the basis of all the information available to them. Second, insofar as the interviewer debriefing questionnaire involves the use of a structured protocol of questions, it was subject to the same types of response-error problems (e.g., satisficing, misinterpretation of question meaning, desirability bias) that affect the instruments it was designed to evaluate. A third weakness, associated specifically with the data compiled from our interviewer debriefing questionnaire, had to do with statistical power. Even when there appeared to be clear preferences among interviewers for a particular question or series of questions, sample sizes were too small to run statistical comparisons. A fourth weakness, associated primarily with focus groups, has to do with the social dynamics of small groups. To be more specific, focus groups are subject to various uncontrolled factors--such as the style/status of the moderator and the personalities of the participants--that can have a significant effect on the nature and quality of the outcome data. For example, in some contexts, the opinions and concerns of quieter participants may not be fully expressed or may be influenced disproportionately by the views of a highly educated/experienced participant.

3. Field-based respondent debriefings. To obtain information about respondents' understanding of CPS questions, we conducted field-based respondent debriefings with household respondents after their fourth and final monthly interview. This post interview consisted of 11 vignettes and 67 follow-up questions. Respondents' eligibility for a set of follow-up questions was determined by their responses during the main interview. To minimize burden, respondents received either all of the vignettes or an average of two or three sets of follow-up questions. Respondent burden was further reduced by only inquiring about one eligible household member for each series of questions.

The follow-up questions used in this CATI/RDD study differed from more typical frame-of-reference probes in three ways (e.g., Cannell et al., 1989; Schuman, 1966). First, the CATI/RDD follow-up questions were more structured than typical cognitive laboratory questions (e.g., the majority were either yes/no questions or open-ended questions with prespecified response categories for field coding). Second, many of the follow-up questions dealt with frame-of-reference issues only indirectly. For example, respondents were not directly asked their definition of a business. Those that did report a household business were asked a series of follow-up questions to determine if the businesses they reported were consistent with the CPS definition. Finally, due to the existence of CATI technology, it was possible to ensure that respondents received follow-up questions based on their specific answers to questions in the main survey.

The CATI/RDD 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 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. Some examples of these uses are provided below.

Follow-up questions that probed for whether a common understanding existed were found to be particularly useful. For example, in the body of the survey, individuals who identify themselves as having more than one job are asked several questions about their main jobs and all other jobs combined. However, what is meant by "main job" is never specified within the context of the survey. In the debriefing, multiple job holders' understanding of the concept "main job" was probed with the following question:

DQ. "You mentioned earlier that you had more than one job. How did you decide which job was your MAIN job?"

Analysis of the responses to this question revealed that respondents did not share a common definition of "main job." Only sixty-three percent of the responses were field-coded as the "job worked at the most hours." Although it was reassuring to find that respondents' understanding of main job was in accord with the core CPS definition (i.e., the job at which the person worked the most hours), the fact that a large minority did not indicates that a definition of this concept should appear in the CPS questionnaire.

An analysis focusing on persons on layoff illustrates the use of follow-up questions to detect whether questions in the main survey are superfluous. In the main CPS survey (versions B and C only), individuals who reported that they were on layoff were asked if they expected to be recalled in the next six months and whether they had been given a date to return to work. There was some speculation that if few recall dates were more than six months away, the question in the main survey regarding a date to return to work could be eliminated. However, a follow-up question revealed that approximately 9% of the recall dates (n=71) were more than six months away, suggesting that the main survey question should be retained.

An evaluation of the accuracy with which casual employment was measured highlights the use of follow-up questions to identify whether certain questionnaire items did a better job of measuring specific activities. Each version of the questionnaire inquired about economic activities during the last week differently:

A. "Did ... do any work at all LAST WEEK, not counting work around the house?"

B. "LAST WEEK, did ... do any work for pay or profit?"

C. "LAST WEEK, did ... do any work at all? Include work for pay or other types of compensation."

To ascertain if one version of the question did a better job of capturing casual employment, individuals who said "no" to these questions were asked in the follow-up questions if they had done any informal work such as babysitting, housepainting, repair work, or bookkeeping. The results indicate that, although version C may do a marginally better job of capturing casual employment, the percentages of missed employment were small for all three versions of the questionnaire (A=2.0%, B=1.8%, and C=1.1%; A/B: X^2=0.09, p=.762; A/C: X^2=1.78, p=.182; B/C: X^2=1.04, p=.308; df=1; n_A=988, n_B=821, n_C=752).

A less standard application of follow-up questions is to use them to construct comparable subsets of respondents from different questionnaire versions without biasing the collection of data within the main survey. An example of this use occurred in the analysis of the hours series. Version B asked individuals a single question about how many hours they worked in the last week. Version C respondents, in contrast, were asked a detailed series of questions: first they were asked about usual hours, then about exceptions that occurred last week, and finally about the actual number of hours worked last week. The version B follow-up questions asked respondents about exceptions to their work schedule last week. Consequently, respondents who had worked extra hours or who had lost hours could be identified in version B for comparison with the same type of respondents in version C. Since the additional questions were part of the post-interview respondent debriefing, hours reported from the shorter version B question in the main survey were not contaminated.

We also tested respondents' understanding of labor force concepts and the consistency of their definitions with official CPS definitions using vignettes. For each of the eleven hypothetical examples of "work" and "looking for work" activities, respondents were asked how they would report a person in that given situation. Results for the vignette data were mixed. Analysis of the four vignettes concerning the classification of marginal types of work (e.g., unpaid work in a family business) showed that, for two of the vignettes, version B respondents were significantly more likely than either version A or C respondents to incorrectly classify marginal types of work as "not working". In contrast, for the two vignettes concerning the classification of nonwork activities (e.g., volunteer services), version B respondents correctly classified both vignettes a significantly larger percentage of the time than did either version A or C respondents. [For a more detailed description of the use of vignettes in respondent debriefing, see Campanelli et al. (1989) and Martin et al. (1991).]

In general, the respondent-debriefing methodology has several strengths associated with it. First, the addition of follow-up questions and vignettes to the conclusion of the main survey provided a relatively cost-effective way to obtain information on respondents' understanding of the questionnaire. Second, attaching follow-up questions and vignettes to a field-based survey enabled us to obtain information from a representative cross section of the population. Third, knowing respondents' associated demographic characteristics made it possible to investigate if certain additional questions needed to be included for certain groups. In addition, we were able to easily identify small groups of individuals (e.g., unpaid family workers), who tend to be hard to find and recruit for interviews in a formal laboratory setting. Fourth, because the follow-up questions were "answer-keyed" on a CATI system, it was possible to obtain information from individuals for whom the follow-up questions were directly relevant. Fifth, as opposed to a laboratory setting, respondents were asked these questions within the context of a survey and an actual interview setting. Finally, the structured format of the follow-up questions provided a relatively standardized way to compare different wordings of a question. In conjunction with the large number of observations, this standardization made it possible to select question wordings based on statistical tests.

Despite these advantages, there are three main weaknesses in using these types of follow-up questions and vignettes. The first is that hypotheses about problem questions or concepts in the main survey have to be established in advance so that appropriate follow-up questions and vignettes can be developed. A second weakness is that follow-up questions and vignettes are subject to their own sources of response error. Further, follow-up questions targeted at specific groups of workers are predicated on respondents correctly understanding the main survey questions used to identify the group. Lastly, it became clear in our work that follow-up questions and vignettes can train and/or sensitize interviewers to what the acceptable responses are.

Item-based response analyses. In addition to the more common approaches of evaluating questions described above, we also examined nonresponse data and response distributions for comparable question sets. The purpose of these analyses was to determine the extent to which different wordings and/or question sequencing produced different distributions of responses. These analyses included interview data for over 72,000 persons.

Item nonresponse rates were defined as the percent of persons eligible for a question who did not provide a substantive response; this included persons who refused to answer and persons who said they did not know the answer. A nonresponse rate was calculated for every question appearing in versions A, B, and C. Separate rates for refusals and don't knows were also examined. Overall, revised question wording did not have a negative effect on item nonresponse. Item nonresponse rates were extremely low (less than 3%) in nearly all questions, and statistically significant differences were found across only a few comparable question sets. [Note: Chi-square values reported in this section account for the clustering within the survey design using a Rao-Scott adjustment procedure (Rao and Scott, 1984).]

Analysis of separate refusal and don't know rates were useful to indicate variation in the extent to which respondents found questions to be sensitive or were able to provide an answer. For example, the refusal rate for the two different versions of the hourly earnings question (see below) differed significantly, 11% vs. 6%, respectively (A/B: $X^2=13.53$, $p<.001$; df=1; $n_A=1504$, $n_B=1300$):

A. "How much does ... earn per hour?"

B. "What is ...'s hourly rate of pay on this job, before deductions of any kind?"

Although, there was a substantial number of "don't knows" (approximately 16%), there was virtually no difference in the don't know rates between these two questions. From these data, we can surmise that the question wording in version A may be more sensitive than that in version B. The phrase "how much" appears to be more offensive to respondents than the phrase "what is ...'s hourly rate of pay."

For the three different versions of the "when did ... last work question" (see below), it was hypothesized that version C would have a higher don't know rate than either version A or B:

A. "When did ... last work at a full time job or business lasting 2 consecutive weeks or more?"

B. "When did ... last work for pay for 2 weeks or longer?"

C. "In what month and year did ... last work for pay for 2 weeks or longer?"

It was hypothesized that the version C question, which requests very specific information (month and year) from the respondent, represents a more difficult recall task than does the question appearing in either version A or version B. This hypothesis was supported. As expected, version C resulted in a significantly higher proportion of don't know responses than either version A or B (7% vs. 2% vs. 2%, respectively; X^2=31.24, p<.001; df=2; n_C=575, n_A=798, n_B=745).

Response distributions, calculated on the basis of eligible persons who responded to the question, were analyzed for those questions which differed between questionnaires. This included questions with different question wording and/or different response categories, as well as cases where one question in the control questionnaire (version A) was divided into two or three questions in the alternate questionnaires (B and C). For some questions which did not differ across versions (e.g., questions on unpaid family workers), the response distributions were useful in determining whether the given question should be retained or whether other questions should be added.

Response distribution analyses can be useful in determining the impact of different design features, such as using direct questions versus indirect questions. For example, in the CPS there is much interest in knowing if the reason a person usually works part time (i.e., less than 35 hours) is due to "economic" reasons (e.g., slack work) versus "noneconomic" reasons (e.g., in school). The two different approaches were used in versions B and C. Individuals who indicate that they want a full-time job are asked:

B. "What is the main reason ... is not working full time? [Interviewers "field code" the response into one of the ten listed response categories which included the two economic reasons "slack work/ business conditions" and "could only find part-time work."]

C.1 "Is the main reason ... is working part time because ... could only find part-time work?"

.2 If "Yes" ask: "Is the main reason because of business conditions or financial problems at ...'s place of employment?"

.3 If "No" ask: "What is the main reason ... is not working full time?"

As can be seen in version C, two direct questions on working part time for economic reasons were asked first, followed by an open-ended question. Relative to B, the use of the direct questions in C produced a much larger--and, we suspect, an erroneously inflated--percentage of "persons wanting full-time work" reported to be working part time for economic reasons (B=40% vs. C=65%; B/C: X^2= 73.26, p<.001; df=1; n_B=645, n_C=521).

Knowing the response distribution for given questions can also be valuable when one is trying to determine if additional questions are useful. For example, the test showed that a direct question about the presence of a family business or farm and a direct question asking non-working family members if they did any unpaid work in the family business or farm were useful. Twenty-two percent of sample persons in households with businesses who were not reported to be working for pay were reported to have done unpaid work in the family business. Without these questions, the work activities of these individuals could have been potentially missed.

The principal strength of response distribution analysis is that it can demonstrate how different patterns of responses can be obtained with different question wording or questionnaire designs. In this respect, the technique was very useful in helping us to select the best questions for inclusion in the version D questionnaire. A weakness associated with response distribution analysis is that it does not necessarily tell us whether one version of a question produces a better understanding of what is being asked. In such situations, the experienced survey designer/researcher is left to rely on inference to explain why response patterns are different. This is precisely why response distribution analyses cannot be a sole method for evaluating modifications in question wording or item sequencing. It is useful only in conjunction with other analytical methods (e.g., respondent and interviewer debriefing).

Discussion

In the first part of this section, we describe how data from the four general methodologies used in phase one of the CATI/RDD pretest were evaluated in developing recommendations for which questions to include in version D, the alternate questionnaire for phase two. The second part of the section is devoted to comparing and contrasting the various methods/techniques used in this pretest.

The general decision-making strategy used to evaluate items in comparable question sets is described below. We initially began by reviewing all of the available methodological data for a particular set of comparable questions from versions A, B, and C. After this review, we drafted a recommendation supporting a particular item from the set. This process was followed in selecting questions for each of the 14 CPS question series (e.g., actual hours, on layoff, earnings). Since the relative usefulness of each of these methods/techniques in selecting the best question was unknown, we did not start out with any *a priori* formulae for weighting methodological data. However, over time, an analytical pattern emerged. We implicitly began to weight response-distribution data and respondent-debriefing data over that obtained from nonresponse analyses, behavior coding, and interviewer debriefing. Our rationale was that response-distribution analyses and respondent-debriefing analyses involved substantially more data/observations *on a larger number of questions* than did other methods/techniques. So while we obtained some very useful information from interviewer debriefing and behavior coding, these analyses simply did not possess either the statistical power or the scope of the other two methods. In most cases, interviewer-debriefing data were consistent with data from more quantitative analyses (e.g., behavior coding, respondent debriefing). Where they were not consistent with these analyses, interviewer-debriefing data were generally overruled. This often took place when it was clear that interviewers were thinking more in terms of ease of question implementation than in terms of other survey factors (e.g., data quality). For example, the B and C versions of the *actual hours* series had very different structures--version B consisted of a single question, while C consisted of a series of five questions. Because there was only one question to ask, interviewers preferred version B; but data from the response-distribution analyses and respondent-debriefing analyses indicated that the longer series of questions (version C) produced more accurate results.

The fact that analytical data produced by diverse methods occasionally point in different directions when focused on a particular question set should not be viewed negatively. As long as there is a consistent pattern of methodological convergence/agreement across comparable question sets, decisions can be made confidently. We believe that plans for pretesting questionnaires should incorporate multiple analytical methods. As noted above and discussed below, each of the methods used in this pretest has a unique set of strengths that either complement or reinforce the strengths of the other methods, and a corresponding set of weaknesses that were essentially negated by the strengths of the other methods.

Our general recommendations, then, can be summarized as follows: (1) use analytical methods that draw on the experiences of interviewers and respondents; (2) document interviewer-respondent interactions within a natural survey context; (3) use response-distribution data as an analytical tool if two or more alternate questionnnaires are being tested, and especially if something is known about "true" distributions or if the questionnaire contains items with sensitive content; and (4) utilize as many of these methods as your organization can reasonably afford--there is nothing more reassuring to a researcher than when diverse methodologies converge on a given conclusion or course of action. Having made the point that all of the methods used in this pretest have intrinsic value, we now provide a more detailed assessment of each method.

Behavior coding is a useful method for identifying problematic questions/series within a particular questionnaire. It is less useful for detecting the effects of subtle differences in question wording when one is evaluating comparable question sets on two or more versions of a given questionnaire. Relative to other methods, behavior coding is labor intensive (e.g., transcribing coded data from monitoring forms for subsequent computer analyses). Implementation of this methodology (i.e., coding live interviews) can also be a demanding task.

The interviewer debriefing techniques (i.e., self-administered questionnaire, focus groups) we used were easy to implement and somewhat less difficult than other methods to develop (e.g., preparation of debriefing questionnaire) and analyze (e.g., content analyses of open-ended questions). The data were useful in corroborating results from other methods (e.g., behavior coding) and in providing explanations for why interviewers and respondents might be having problems with specific questions or series. Although basically qualitative in nature, even a few focus groups can be valuable for providing insights into problems and suggesting possible solutions.

The respondent debriefing method represents a very useful analytical tool. The follow-up questions, in particular, made it possible to detect problems that response-distribution analyses and the other methods did not reveal. Follow-up questions and vignettes were useful for helping researchers to understand why certain questions may be posing conceptual problems for respondents. For example, none of the other methods revealed any problems with respondents' comprehension of the phrase "main job." However, analysis of the responses to a follow-up question revealed that respondents did not share a common understanding. Respondent-debriefing techniques can also be used productively to determine if respondents have narrower or broader interpretations of the concepts being measured than the survey designer intends. The technique of using follow-up questions, in particular, is very useful in helping researchers to select the best questions among comparable question sets. As one might expect, given their diagnostic value, respondent-debriefing techniques require that a substantial amount of time be set aside for preparatory work (e.g., developing a set of item-specific debriefing questions and concept-related vignettes) and for analyses. Further, the usefulness of this technique depends a lot on the skill of the survey researcher in anticipating where problems will crop up and in designing effective debriefing questions (Campanelli et al., 1991; Cannell et al., 1989). One of the more practical advantages of this method is that it is possible to include debriefing questions at the end of a standardized survey without substantially increasing survey length and without elaborate interviewer training measures.

Item-based response analysis (i.e., nonresponse and response-distribution analyses) is a unique methodology in that once the questionnaire is developed, the preparatory work only involves the time and effort is required to conduct these analyses (e.g., writing computer programs to screen or to analyze item-specific data). The utility of this technique is in indicating where patterns of response are different for large groups of respondents. A shortcoming of this technique is that it does not identify the "correct" response pattern, unless the researcher already knows from another source what the correct pattern should look like. Nonresponse data are useful for identifying the most/least objectionable question in a comparable question set and for evaluating the sensitivity of single questions. We recommend its general use even though its utility for us was limited (i.e., item nonresponse rates for the CPS are very low relative to other surveys).

To conclude, we encourage further research comparing and contrasting these methods/techniques so that a framework for pretesting questionnaires can be established and a set of guidelines developed for choosing among techniques when only limited resources are available. We also would like to see these methods/techniques used for evaluating other questionnaires, so that the reliability of the data produced by these techniques can be established.

Acknowledgements

We would like to thank Jennifer Hess, Steve Miller, Laura Spencer, Athar Malik, and Sue Zimmerman for their invaluable technical assistance. We would also like to express our appreciation to other members of the BLS/Census Questionnaire Development Workgroup who played an active role in all aspects of the CPS CATI/RDD test (Sharon Cohany, Kathy Creighton, Harvey Hamel, Howard Hayghe, Jonathan Kurlander, and Paul Mullin).

References

Belson, W.R. (1981). The Design and Understanding of Survey Question. Aldershot, England: Gower.

Bradburn, N., Sudman, S. and Associates. (1979). Improving Interview Method and Questionnaire Design. San Francisco: Jossey Bass.

Brenner, M. (1982). Response effects of 'role-restricted' characteristics of the interviewer. In W. Dijkstra and H. Van der Zouwen (eds.), Response Behavior in the Survey Interview. London: Academic Press.

Bureau of Labor Statistics. (1988). Response Errors on Labor Force Questions: Based on Consultations with Current Population Survey Interviewers in the United States. Paper presented at the meeting of the OECD Working Party on Employment and Unemployment Statistics, Paris, France.

Campanelli, P.C., Martin, E.A., and Creighton, K.P. (1989). Respondents' understanding of labor force concepts: Insights from debriefing studies. Proceeding of the Census Bureau's Fifth Annual Research Conference. Washington, DC: Bureau of the Census, 361-374.

Campanelli, P.C., Martin, E.A., and Rothgeb, J.M. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. The Statistician, 40, 253-264.

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

Cannell, C., Oksenberg, L., Fowler, F.J., Kalton, G., and Bischoping, K. (1989). New Techniques for Pretesting Survey Questions, Final Report for Grant Number HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment. Ann Arbor, MI: Survey Research Center, University of Michigan.

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

Converse, J.M., and Presser, S. (1986). Survey Questions: Handcrafting the Standardized Questionnaire. Newbury Park, CA: Sage Publications.

Converse, J.M. and Schuman, H. (1974) . Conversations at Random: Survey Reasearch as Interviewers See It. New York: John Wiley.

Copeland, K. and Rothgeb, J. (1990). Testing Alternative Questionnaires for the Current Population Survey. In American Statistical Association, Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

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

Dijstra, W., Ver der Veen, L., and Van der Zouwen, J. (1985). A field experiment on interviewer-respondent interaction. Chapter 3 in Brenner, et al. (eds.), The Research Interview. London: Academic Press.

Ferber, R. (1956). The effect of respondent ignorance on survey results. Journal of the American Statistical Association, 51 (276), 576-586.

Fracasso, M.P. (1989). Reliability and validity of response categories for open-ended questions in the current population survey. In Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

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

Marquis, K. (1969). Interviewer-Respondent Interaction in a House hold Interview. Paper presented at the annual meeting of the American Statistical Association, New York City.

Marquis, K. (1971). Effects of race, residence, and selection of respondent on the conduct of the interview. In J. Lansing, S. Withey, and A. Wolfe (eds.), Working Papers on Survey Research in Poverty Areas. Ann Arbor, MI: Institute for Social Research.

Martin, E.A., Groves, R., Mattlin, J., and Miller, C. (1986). Report on the Development of Alternative Screening Procedures for the National Crime Survey. Washington, DC: Bureau of Social Science Research.

Martin, E.A., Campanelli, P.C., and Fay, R.E. (1991). An application of Rasch analysis to Questionnaire Design. The Statistician, 40, 265-276.

Mathiowetz, N.A., and Cannell, C.F. (1980). Coding Interviewer Behavior as a Method of Evaluating Performance. In American Statistical Association, Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association, 525-528.

Morton-Williams, J. (1979). The use of 'verbal interaction coding' for evaluating a questionnaire. Quality and Quantity, 13, 59-75.

Nelson, D. (1985). Informal testing as a means of questionnaire development. Journal of Offical Statistics, 1(2), 179-188.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New strategies for pretesting survey questions. Journal of Official Statistics, 7(3), 349-365.

Oksenberg, L. (1981). Analysis of Monitored Telephone Interviews. Research Report to Bureau of the Census (JSA 80-23). Ann Arbor: Institute for Social Research.

Palmisano, M. (1989). Respondent Understanding of Key Labor Force Concepts Used in the CPS. In American Statistical Association, Proceding s of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Presser, S. and Zhao, S. (1990). Attributes of Questions and Interviewers as Determinants of Interviewing Performance. Unpublished paper. College Park, MD: University of Maryland.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. Annals of Statistics, 12, 46-60.

Rothgeb, J.M., Polivka, A.E., Creighton, K.P., and Cohany, S.R. (1991). Development of the proposed revised Current Population Survey. Paper presented at the annual meeting of the American Statistical Association, Atlanta, GA.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. American Sociological Review, 31, 218-222.

Shepard, J.D. and Vincent, C.J. (1991). Interviewer-respondent interactions in CATI interviews. Proceedings of the Census Bureau's 1991 Annual Research Conference, forthcoming. Washington, DC: Bureau of the Census.

Tanur, J.M. and Fienberg, S.E. (1990). Cognitive aspects of surveys: Yesterday, today and tomorrow. Paper presented at the International Conference on Measurement Errors in Surveys, Tucson, AZ.