THE USE OF INTERVIEWER DEBRIEFINGS TO IDENTIFY PROBLEMATIC QUESTIONS ON ALTERNATIVE QUESTIONNAIRES

James L. Esposito, Bureau of Labor Statistics,

and

Jennifer Hess, Bureau of the Census

Paper prepared for presentation at the annual meeting of the

American Association for Public Opinion Research

St. Petersburg, FL

May 1992

This paper reports research undertaken by staff at the Bureau of Labor Statistics (BLS) and the Bureau of the Census. The views expressed are attributable to the authors and do not necessarily reflect the views of either BLS or the Census Bureau.

INTRODUCTION

There would appear to be general consensus in the survey methodology literature that pretesting a questionnaire is an integral part of the questionnaire development process (e.g., Converse and Presser, 1986; DeMaio, 1983a; Nelson, 1985). And while there are a variety of methodologies and techniques available for pretesting questionnaires (Bercini, 1991; Cannell, Oksenberg, Kalton, Bischoping, and Fowler, 1989; DeMaio, 1983a; Willis, Royston, and Bercini, 1991), what constitutes a useful pretest methodology/technique is not clear. In a recent work on this general topic, Converse and Presser (1986) assert:

> "There are no general principles of good pretesting, no systemization of practice, no consensus about expectations, and we rarely leave records for each other. How a pretest was conducted, what investigators learned from it, how they redesigned their questionnaire on the basis of it--these matters are reported only sketchily in research reports, if at all. (p. 52)"

Recently, however, there has been a concerted effort among survey researchers to demonstrate the utility--and oftentimes the relative utility--of various pretesting methodologies (e.g., Cannell, et al. 1989; Campanelli, Martin, and Rothgeb, 1991; Blair and Presser, forthcoming; DeMaio, 1983a; Oksenberg, Cannell, and Kalton, 1991; Sykes and Morton-Williams, 1987; Willis, 1991; cf. Esposito, Campanelli, Rothgeb, and Polivka, 1991). Most of these efforts have focused on newer methodologies (e.g., behavior coding, respondent debriefing using structured probes) and, in so doing, have given short shrift to the traditional and most widely used method of questionnaire

pretesting, interviewer debriefings (for an exception, see DeMaio, 1983b). In the present paper, we attempt to correct this imbalance by reviewing very briefly what survey methodologists have had to say recently about interviewer debriefings, by describing our experiences with this methodology, by identifying some of the strong points and limitations of the interviewer-debriefing techniques we used, and by advocating a more integrative approach to debriefing interviewers.

Interviewers are in a unique position to evaluate the merits of survey questions (Converse and Schuman, 1974; DeMaio, 1983b). Not only do they obtain very useful feedback from respondents in the course of administering questionnaires, more experienced interviewers can often draw on their accumulated knowledge of survey interactions to identify--during the pretesting stage of questionnaire development--questions that are likely to cause problems for interviewers and respondents. In her discussion of interviewer debriefings (i.e., individual and group debriefings) and structured post-interview evaluations (e.g., ratings), DeMaio (1983b) provides several examples from pretesting work with large governmental surveys that appear to demonstrate the utility of these techniques. In her discussion of one survey, contributions that interviewers made in improving the questionnaire's overall design were grouped into four categories: question wording, question sequencing, reference periods, and format and physical features of the questionnaire.

DeMaio also compared information collected from oral interviewer debriefings and written post-interview evaluation forms and found that the two methods provide complementary data. Relative to oral debriefings, evaluation forms provided "a more exact enumeration" of problematic questionnaire items and corresponding estimates of the number of interviewers who experienced problems with those items (i.e., prevalance estimates). Relative to evaluation forms, the oral debriefing format provided information that in certain respects was qualitatively different. For example, in group discussions, interviewers often went beyond simply identifying a questionnaire item as problematic and suggested possible reasons and solutions for the problem; also, they sometimes expressed concerns about data quality (e.g., underreporting of sensitive data) that was not specific to a particular item (see DeMaio, 1983b, for a more detailed discussion).

In contrast, recent methodological research suggests that traditional pretesting methods (i.e., interviewer debriefings) may be deficient in a number of ways. Bischoping (1989) identifies and discusses four problem areas. The first has to do with the completeness of reports regarding problematic questions. Simply put, in group debriefing sessions, interviewers rarely mention *all* of the problems encountered when administering the questionnaire. A second area of concern has to do with estimating the prevalence of problems. For example, when a problem with a particular question is identified by one

interviewer, there is generally no attempt by the moderator to find out how many other interviewers experience the same problem. When interviewers do corroborate that a question is problematic, they do not always agree as to what the problem is. And when judgments of prevalence are offered more or less spontaneously by interviewers, the quantifiers used tend to be vague or ambiguous (e.g., "*most/some* respondents think the question is too sensitive"); estimates of this sort are of dubious value to researchers. A third area of concern has to do with the accuracy of reports of interviewer experiences. For example, sometimes when interviewers identify a particular question as problematic (e.g., difficult to read as worded), that very same question is not flagged as problematic when another evaluation methodology is used (e.g., behavior coding). A final area of concern has to do with the reliability of interviewer debriefings. Here the issue is agreement between independent groups of interviewers as to whether--and why--a particular problem exists (e.g., unclear instructions, item sensitivity). For some types of problems (e.g., difficulty reading the question as worded), reliability/ kappa statistics are quite low.

Though somewhat harsh, Bischoping's critique of the interviwer debriefings methodology has merit. Two recent studies (Fowler, 1989; Willis, 1991) provide empirical support for her contention that interviewer debriefings *sometimes* miss problems identified by other pretesting methodologies (i.e., behavior coding). And with regard to the issue of prevalence, DeMaio (1983b, p. 120)

has also pointed out that one of the weaknesss of individual and group debriefings is that they will not yield prevalence data for problematic questions. But the evidence is not all negative. Both studies alluded to above demonstrate that there is considerable overlap between behavior coding and interviewer debriefings in terms of identifying problematic survey questions. Willis (1991), for example, reports evaluative agreement on 113 of 152 survey items (kappa=.5; concordance rate=74%, $X^2(1)=42.3$, $p<.001$) using behavior coding and interviewer debriefing methods. So, from our perspective, the issue is not the utility of interviewer debriefings--or even whether survey methodologists should use interviewer-debriefing techniques for pretesting questionnaires now that other methods are available (e.g., behavior coding, respondent debriefings). We believe that debriefing interviewers is a very useful pretesting methodology and we hope to provide evidence in support of that belief.

RESEARCH CONTEXT

As part of the overall effort to redesign the Current Population Survey (CPS) questionnaire, the Census Bureau--in collaboration with the Bureau of Labor Statistics--is conducting a multi-phase field test of alternative CPS questionnaires. The first two phases of this field test utilized Computer-Assisted Telephone Interviewing (CATI) and a Random Digit Dialing (RDD) sampling plan, and hence is referred to as the *CATI/RDD Test*. Phase one of the CATI/RDD test (July 1990 to January 1991) involved

approximately 72,000 interviews; its purpose was to compare the current version of the CPS questionnaire ("A") with two 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 this first phase was a single alternative questionnaire ("D"), which comprised the best questions from versions A, B, and C, as well as any questions deemed necessary given the results of phase-one analyses. The second phase of the CATI/RDD test (July to October 1991) involved approximately 30,000 interviews. During this phase, the current CPS questionnaire (A) was tested against the alternative questionnaire (D) produced in phase one. The purpose of phase two was to fine tune version D. In July 1992, we will begin the final phase of the redesign: the *CATI/CAPI Overlap Test*. The overlap test, which will take 18 months to complete, "will be used to estimate the combined effect of the new questionnaire [D'] and the use of CATI/CAPI on the labor force estimates" (Copeland and Rothgeb, 1990).

METHODS

Phase One. During phase one of the CATI/RDD test, two techniques for debriefing interviewers were used: (1) completion of a self-administered debriefing questionnaire, and (2) active participation in a focus group with other interviewers. Though the two aspects of interviewer debriefing utilized different formats, they sought to collect similar information and, as a

result, shared a similar underlying structure. The debriefing instruments (i.e., the questionnaire and the focus-group guidelines) were structured to proceed from general preferences regarding questionnaire versions (e.g., which version flowed the best/worst) to specific evaluations of a particular question or series of questions (e.g., which question or series of questions was most difficult to ask).

Interviewer-Debriefing Questionnaire. Each CATI interviewer was asked to complete a self-administered debriefing questionnaire. The questionnaire was distributed about ten weeks after the beginning of phase one (September 1990) and was completed by 88 percent (68 of 77) of the interviewers who participated in this phase. The questionnaire was administered prior to the focus groups so that answers to debriefing questions would not be influenced by focus group discussions.

To measure general impressions during phase one of the CATI/RDD test, interviewers were asked which of the three CPS questionnaires they liked the most, which flowed the best/worst, and why. Interviewers were then asked which of the 14 series of questions they found most difficult to ask as interviewers and which they thought was most difficult for respondents to answer; in addition, they were asked to tell us why they believed the series was problematic. Specific questions were then asked regarding which single question was most difficult to ask as an interviewer and which single question was most difficult for

respondents to answer, followed by two additional questions inquiring why they thought the question was difficult to ask/answer and how they would change the question to resolve the problem. In addition, interviewers were asked which concepts or terms they felt respondents most commonly misunderstood or misinterpreted, which question respondents refused to answer most frequently, and the kinds of problems they encountered with proxy and self respondents.

Due to the fact that several questions on the debriefing questionnaire were open-ended [e.g., "Why did you find this question difficult to ask?"; "How would you fix the question to make ot work better?"), response categories had to be developed for coding purposes. Since there were only 68 debriefing questionnaires to evaluate, all of the responses to a particular open-ended debriefing question were transcribed and crudely grouped before creating general coding categories for that question; in some cases, the coding categories generated for one debriefing question could be used for other questions. After the response-coding categories were developed, all 68 debriefing questionnaires were manually coded and, later, these coded data were entered into the interviewer-debriefing database.

Focus Groups. During phase one of the CPS redesign effort, six focus-group sessions were conducted at the Census CATI facility in Hagerstown, Maryland (September through November, 1990). Two focus groups were conducted each month, with 8 to 10 interviewers

per group; each session lasted about two hours. The interviewers who participated in the six focus groups were drawn from the same CATI staff that had completed the interviewer-debriefing questionnaire in early September. Three researchers from BLS and Census served as moderators. Focus-group guidelines and questions were developed prior to the first session and had the effect of standardizing the manner in which the six sessions were conducted.

Phase Two. In contrast to the purpose of phase one (i.e., to select the best questions from alternative CPS questionnaires), the purpose of phase two was to fine tune the alternative CPS questionnaire (version D) developed on the basis of analyses of phase-one data. Insofar as only minor changes to version D were expected as a result of phase two analyses, researchers could focus most of their efforts on version D with the understanding that this would be the revised CPS questionnaire for the 1990s. Given the limited objective of phase two, the only technique used for debriefing interviewers during this phase was focus groups.

Focus Groups. Three focus groups were conducted in September 1991, with each session again involving 8-10 CATI interviewers. Focus-group questions were developed to take advantage of information gained during phase one and to assess the impact of of decisions made in the design of version D. In addition to the more general debriefing questions (e.g., version preference), focus-group items in phase two targeted important CPS questions

(e.g., job search) and important CPS series (e.g., economic part time, industry and occupation) that had been identified as problematic in phase one--and later modified. There was also an important procedural change introduced in phase two: When one interviewer mentioned a problem with a particular questionnaire item or series, moderators tried to assess the extent to which there was general consensus within the group by asking other participants if they also experienced that problem. The intent here was to get an idea of how serious particular problems were and, in so doing, obtain a crude measure of prevalence.

RESULTS

Before providing an illustrative sampling of results from the interviewer debriefings conducted during the first two phases of the CPS redesign effort, we wish to say a few words about the function of interviewer debriefings and the nature of debriefing data. As we understand it, the primary function of interviewer debriefings is to find out what kinds of problems interviewers experience--or observe--while administering a questionnaire. A secondary function is to identify potential ways of resolving those problems. How well these functions are satisfied, some individuals believe, depends on the nature of the data produced by specific debriefing techniques. The information or data collected from interviewers in the debriefing process can be characterized as existing along a qualitative/quantitative continuum. Generally speaking, the information gathered from interviewers during focus groups is qualitative in nature; these

data are verbal, subjective, non-numerical and, therefore, not statistically analyzable (see Box 1). The information gathered from interviewers via debriefing questionnaires generally produce information that is quantitative in nature; these data are often expressed in numerical terms (e.g., simple counts, percentages, response distributions), are sometimes amenable to statistical tests of significance, and are generally viewed as less subjective relative to focus-group data (see Box 2). Most analysts would probably agree that quantitative data are preferable to qualitative data when resources are plentiful (e.g., money, the amount of time and effort interviewers are willing to expend on evaluation tasks). When such resources are not plentiful, researchers should take full advantage of qualitative data. Though we debriefed interviewers using both focus groups and self-administered questionnaires, most of the debriefing data we collected was either qualitative or non-statistically quantitative (i.e., simple counts).

Illustrative Results. Given the opportunity to speak their minds, there are a number of areas where interviewers provide very useful information regarding questionnaire or item-specific problems. Some of the more important contributions made by interviewers in the course of our debriefings are summarized below.

BOX 1: Qualitative Data (Focus Group)

Q1. Of the three CPS questionnaires--the current CPS (version A), version B, or version C--which do you like most and why?

INT 1: A, because questions are phrased better overall; questions are plainer and more understandable.
INT 2: B; like the specific terms that are used in B, especially when you have a person who has more than one job.
INT 3: B; like the dependent industry and occupation (I/O) questions, especially the one that has to do with the respondent's occupation.

BOX 2: Quantitative Data (Debriefing Questionnaire)

Q1. Of the three CPS questionnaires--the current CPS (version A), version B, or version C--which do you like most and why?

Distribution	A	B	C
Number (N=68)	13	43	12
Percent[1]	19%	63%	17%
Reasons[2]			
Easier to understand	3	5	2
Worded better	1	7	3
Flows the best	1	9	3
Q'aire shorter/ more concise	4	9	1
More direct	7	9	2
Dependent I/O	0	14	1
Shorter questions	2	4	1
Less burdensome	0	2	1
Other	3	6	3

[1] Percentages are significantly different from one another [$X^2(2)=27.0$, $p<.005$].
[2] The number of reasons for liking a particular questionnaire may be greater than the number of interviewers who chose that version, because some interviewers provided more than one reason for liking that particular version.

1. *Problematic Concepts/Terms*. Interviewers are particularly effective at identifying concepts/terms that they or respondents have difficulty understanding. Some of the concepts/terms identified as problematic in our debriefings were: "profit", "compensation", "private company", "union or employee association contract", "owning vs. operating a business", and "main job". We were surprised to learn, for example, that a word as ordinary as "profit" was causing problems for respondents. As it turned out, the problem is not simply with the word itself, but with the context in which the word is embedded (see item 3). Another term that was problematic for some respondents who were multiple job holders was the concept of "main job" (e.g., "How many hours per week do you USUALLY work at your *main job*?"). The problem here is that there are a variety of ways a worker can define *main job*: job worked the most hours (official BLS definition), job that pays the most, job worked at the longest. Respondent debriefing analyses revealed that 63% of multiple job holders define *main job* in a manner consistent with the official definition; the other 37% were using a different definition. The solution here was to include the definition of *main job* in the body of the question (i.e., "How many hours per week do you USUALLY work at your *main job*? By main job we mean the one at which you usually work the most hours.").

2. *Problems with the Structure of Questions*. Interviewers are also adept at identifying structural problems with particular questions; this type of problem is often totally transparent to

survey designers and researchers. Consider the following question: "Does anyone in this household have a business or farm?" Seems straightforward enough; however, in one focus group, interviewers told us that some respondents misunderstand this question and appear to hear: "Does anyone in this household have a business or *firm*?" It makes sense; interviewers tend to read questions quickly and respondents may hear *firm* as a synonym for the word *business*. In addition to detecting the problem, interviewers also provided us with a very simple solution--add the article, "a", before the word "farm" (i.e., "Does anyone in this household have a business or *a* farm?") and tell interviewers to slow down when reading the question.

3. *Problems Attributable to Question Sequencing*. Interviewers can also help to identify problems attributable to question sequencing. For example, the first labor force question on one of the alternative questionnaires tested in phase one asked: "Do you or anyone in this household have your own business or farm?" The next question asked: "LAST WEEK, did you do any work for pay or profit? As mentioned in item 1 above, the term "profit" appeared to cause problems for quite a few respondents. Some respondents would answer, "No, but I did have a job." It would appear that the concept of "profit" is simply not relevant for the vast majority of workers who do not own a business; further, these two questions--considered as a unit--apparently create an expectation in some respondents' minds that "these questions are requesting information about businesses". As a potential

solution to the problem, a few interviewers suggested that we drop the word "profit" from the second question; but that would cause problems for persons who owned a business and worked for profits--not for a paycheck. The solution ultimately adopted was to reword the *work* question as follows: "LAST WEEK, did you do ANY work for (*either*) pay (*or profit*)? The italicized words were included in the question *only if* the specified person was identified in the prior question as having a business or a farm.

4. *Problems with a Particular Type or Class of Questions.* Interviewers usually will not hesitate to tell researchers when they, or respondents, are experiencing problems with a particular type or class of questions, though it is difficult to tell sometimes with whom the problem actually lies. It was very clear from our focus groups and questionnaire data that interviewers and respondents struggled with the earnings questions. Data compiled from the phase one debriefing questionnaires illustrates where the problem with this question series lie (see Box 3). Most of the interviewers who identified this series as the one that was most difficult for them to ask--and for respondents to answer--noted that it was the personal nature of the questions (i.e., item sensitivity) that caused the most problems. Even though the questions varied in content across the three versions of the CPS questionnaire, no one version appeared preferable. Interviewers suggested various ways of making these questions less sensitive. Some solutions appeared very reasonable and thus were incorporated into the revised CPS questionnaire.

BOX 3: Earnings Series Data (Debriefing Questionnaire)

Q. Which particular series of questions do you find most difficult to ask as an interviewer (specify version)?
A. **Earnings Series**[1]

Q. Why do you find this series difficult?
A. **Reason**[2] (see below):

	A	B	C
Confusing	2	1	1
Personal	5	3	6
Wordy	1	3	3
Response difficulty	1	2	1
Other	0	0	1

Q. Which particular series of questions do you think is most difficult for respondents to answer (specify version)?
A. **Earnings Series**[1]

Q. Why do you suppose this series poses difficulties for respondents?
A. **Reason**[2] (see below):

	A	B	C
Confusing	2	0	2
Personal	7	8	8
Wordy	0	1	1
Response difficulty	3	3	4

[1] Interviewers (N=68) had a total of 14 question series to choose from on this debriefing question, and were supposed to select one series only.
[2] Column totals may be less than the actual number of interviewers who selected this series, because some interviewers gave more than one reason.

For example, interviewers suggested that a statement be read to hesitant respondents explaining why these earnings questions were being asked in a labor-force survey. The following statement now appears in the revised questionnaire: "READ IF NECESSARY: We use this information to compare the amount that people earn in

different types of jobs." Some solutions (e.g., collecting earnings data using income ranges rather than discrete amounts), though reasonable in principle, would compromise the quality of the survey data obtained. Other solutions were not reasonable (e.g., eliminate the question/series).

It is important to recognize that interviewers are not always impartial evaluators of the survey questions they are required to ask respondents. Sometimes, their evaluations and preferences can lead researchers astray if acted upon without reviewing other analytical data. A good example of this occurred in phase one with the "actual hours" series (see Box 4). Possibly because it involved asking a single question, many interviewers preferred the version B series over the version C series--which could involve asking the respondent as many as five questions. Relative to the B series, the C series was characterized by many interviewers as repetitive and wordy. If it was put to a vote among interviewers, the revised CPS questionnaire would probably have the single actual-hours question from version B. Fortunately, other analyses (i.e., respondent debriefing and response distribution) indicated that the C version of the actual-hours series produced more accurate data, and so this is the series that appears in the revised questionnaire.

BOX 4: Actual Hours Series (Versions B and C)

Version B Question:

Q1. Taking into account any extra hours worked or time taken off last week, how many hours did you ACTUALLY work at your job?

Version C Questions:

Q1. LAST WEEK, did you lose or take off any hours from work for any reason such as illness, vacation, holiday, labor dispute or layoff?
[If "yes", ask Q2; if "no", skip to Q3.]

Q2. How many hours did you take off?

Q3. LAST WEEK, did you work any overtime or extra hours that you do not usually work?
[If "yes", ask Q4; if "no", skip to Q5.]

Q4. How many ADDITIONAL hours did you work?

Q5. So, for LAST WEEK, how many hours did you actually work at your job?

DISCUSSION

As we hope the previous sampling of results demonstrate, the principal strength of the interviewer debriefings used in the CPS redesign was that they enabled researchers to identify (and, when feasible, correct) problems with misunderstood concepts/terms, structural problems with specific questions, problems with question sequencing, and problems with a particular class or type of questions. And, relative to other methodologies, the techniques used were relatively easy to administer--although not always easier to compile and organize for analysis purposes.

Another positive feature of interviewer debriefings, when used as one component of a multi-methodology analysis plan, is that information collected via focus groups and structured debriefing questionnaires complements the more quantitative data obtained from other pretesting methodologies, like behavior coding and response-distribution analyses. Information supplied by interviewers often helps to provide explanations for the patterns observed in quantitative data. However, as Bischoping (1989) has observed, while interviewer debriefings may help to identify some problems with questionnaire items, these techniques provide little information as to the prevalence or magnitude of such problems.

We would agree that the prevalence problem is a serious weakness of the interviewer debriefing methodology as it has been applied in past pretesting work; but it is not an inherent limitation of this methodology. The prevalence problem can be resolved, we believe, by adopting an integrated approach to interviewer debriefings. Before describing such an approach however, we need to make an important connection with a question-asking paradigm suggested by Schuman and Scott (1987). According to Schuman and Scott, in order to understand what any public has "in mind" regarding a particular issue/topic, researchers should: first, obtain spontaneous, free-response expressions by the public on that particular issue/topic; and then use this information to construct a set of fixed-alternative questions with which to follow up. In their words:

> "[When answering closed questions], respondents tend to choose among the alternatives offered to them, even where they are explicitly instructed that this is not necessary. If an investigator wishes to know how the public ranks all alternatives that come to mind, the initial ranking must be provided in a free answer situation. . . . [I]t is possible to proceed in a two-step sequence: first, obtain spontaneous expressions by the public, then use these to construct a set of closed choices." (Schuman and Scott, 1987, p. 958)

We believe this paradigm can be applied to an approach for improving the interviewer debriefing methodology. To do so, we must first define the term "public", and clarify the phrase "in mind". For our purposes, the *public* of interest is the universe of interviewers asking the questions for a specific survey. And the phrase *in mind* refers to the opinions those interviewers have regarding the questionnaire and the questionnaire items they have been working with. The integrative approach that we wish to propose consists of five steps and involves multiple debriefing techniques (e.g., focus groups, debriefing questionnaires). The five step process is outlined below:

Step 1: After the target questionnaire has been administered a sufficient number of times, thoroughly debrief a *sample* of interviewers via focus groups, a structured questionnaire, and/or one-on-one interviews using predominantly open-ended questions. When the interviewer staff is small (e.g., less than 20), we would suggest debriefing the entire staff.

Step 2: Consolidate and categorize the debriefing information collected above. For specific target questionnaire items (and/or issues), identify the most common problems and generate a limited number of categories; some or all of these categories will be adapted for used as response options for specific debriefing-questionnaire items.

Step 3: Develop an interviewer debriefing questionnaire with predominately closed-ended questions and distribute to *all* interviewers.

Step 4: Analyze the data (e.g., based on questionnaire returns, assess the prevalence rates for problematic target questionnaire items).

Step 5: Compare and contrast findings with those generated by other pretesting methodologies (e.g., behavior coding, respondent debriefings). When methodological findings are discrepant, conduct a limited number of followup debriefings (e.g., fifteen one-on-one interviews or two focus groups) with members of the interviewer staff to explore possible reasons for the discrepancy.

We believe such an approach will improve the quality and utility of interviewer debriefings and, in so doing, effectively address many of the criticisms that have been raised with regard to this methodology (e.g., Bischoping, 1989). We should add that at the foundations of this approach lies a fundamental belief that questionnaire pretesting should involve not only multiple techniques (e.g., focus groups, questionnaires, one-on-one interviews), but also multiple methodologies (e.g., interviewer and respondent debriefings, behavior coding). A pretesting plan that relies on one or two methodologies is more apt to miss problematic questionnaire items than one based on three or more methodologies (see Esposito et al., 1991).

REFERENCES

Bercini, D. (1991). Techniques for evaluating the questionnaire draft. Statistical Policy Working Paper 20: Seminar on Quality of Federal Data. Federal Commission on Statistical Methodology, 340-348.

Bischoping, K. (1989). An evaluation of interviewer debriefing in survey pretests. In C. Cannell, L. Oksenberg, G. Kalton, K. Bischoping, and F.J. Fowler (eds.), *New techniques for pretesting survey questions* (Chapter 2). Research Report. Ann Arbor, MI: Survey Research Center, University of Michigan.

Blair, J. and Presser, S. (forthcoming). An experimental comparison of alternative pretest techniques: A note on preliminary findings. *Proceedings of the 9th Annual Advertising Research Foundation Research Quality Workshop.*

Bureau of Labor Statistics, US Department of Labor (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. *Proceedings 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., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F.J. (eds.) (1989). *New techniques for pretesting survey questions.* Research Report. Ann Arbor, MI: Survey Research Center, University of Michigan.

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

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

Copeland, K. and Rothgeb, J. (1990). Testing Alternative Questionnaires for the Current Population Survey. In In the American Statistical Association's *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, 63-71.

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

DeMaio, T.J. (1983b). Learning from interviewers. In T.J. DeMaio (ed.), *Approaches to Developing Questionnaires* (Chapter 9), Statistical Policy Working Paper 10. Washington, D.C.: Office of Management and Budget.

Esposito, J.L., Campanelli, P.C., Rothgeb, J.M., and Polivka, A.E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. In the American Statistical Association's *Proceedings of the Section of Survey Methods Research*. Alexandria, VA: American Statistical Association, 46-55.

Fowler, F.J. (1989). The significance of unclear questions. In C. Cannell, L. Oksenberg, G. Kalton, K. Bischoping, and F.J. Fowler (eds.), *New techniques for pretesting survey questions* (Chapter 5). Research Report. Ann Arbor, MI: Survey Research Center, University of Michigan.

Fracasso, M.P. (1989). Categorization of responses to open-ended labor force questions in the Current Population Survey. In the American Statistical Association's *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, 481-485.

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, 349-365.

Palmisano, M. (1989). Respondent understanding of key labor force concepts used in the CPS. Paper presented at the annual meeting of the American Statistical Association, Washington, DC.

Schuman, H. and Scott, J. (1987). Problems in the use of survey questions to measure public opinion. Science, *236*, 957-959.

Sykes, W. and Morton-Williams, J. (1987). Evaluating survey Questions. *Journal of Offical Statistics*, *3*(2), 191-207.

Willis, G.B. (May 1991). 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, AZ.

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