CHAPTER 24

Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment

James L. Esposito
U.S. Bureau of Labor Statistics

Jennifer M. Rothgeb
U.S. Bureau of Census

24.1 INTRODUCTION

The intent of this chapter, in general terms, is to describe the problem-solving behavior of survey researchers who engage themselves in efforts to detect and minimize sources of measurement error (Biemer *et al.*, 1991; Groves, 1987, 1989; Groves *et al.*, 1988; Turner and Martin, 1984). We are concerned specifically with efforts by survey researchers to obtain high-quality survey data through improvements in questionnaire evaluation and design. As noted by Dippo (Chapter 20), a first step towards implementing continuous process improvement is to define appropriate measures and build them into the survey measurement system. As survey researchers embrace the tenets of the Cognitive Aspects of Survey Measurement (CASM) movement (Jabine *et al.*, 1984; Jobe and Mingay, 1991), they have developed models of the question asking and answering process using concepts from cognitive psychology (Tourangeau, 1984) and social psychology (Cannell *et al.*, 1981; Esposito and Jobe, 1991), and have developed methods and techniques for assessing measurement error. By providing examples and an illustration of how some of these techniques have been used to evaluate survey questions, we hope to convince readers that their use leads to improve-

Survey Measurement and Process Quality, Edited by Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin.
ISBN 0-471-16559-X © 1997 John Wiley & Sons, Inc.

ments in data quality and that their utility extends beyond pretesting to post implementation quality assessment as well. Our emphasis is on the *process* of evaluation, and not the strengths and weaknesses of specific techniques *per se*. (For a discussion of strengths and weaknesses of these methods see Presser and Blair, 1994; Esposito *et al.*, 1992; Cannell *et al.*, 1989; Fowler and Roman, 1992.)

The chapter is organized as follows. In Section 24.1, after briefly addressing the issue of survey quality, we identify and discuss some of the techniques that have been developed to evaluate the effectiveness of survey questions and review how others have used these techniques to assess the quality of data obtained from interviewer administered questionnaires. In Section 24.2, we report on how some of these techniques were used to identify problems and assess improvements in the quality of data from the redesigned U.S. Current Population Survey (CPS), and in Section 24.3, we propose an idealized quality assessment program for major social and economic surveys, address pragmatic issues associated with such a program, and discuss some of the benefits and costs (nonmonetary) associated with quality assessment research.

24.1.1 Survey Quality

Generating a definition of survey quality that researchers from a variety of survey-related disciplines would find acceptable may well be an impossible task. The lack of an inclusive definition, however, does not mean that the concept is poorly understood. Bailar (1984) views survey quality as a multidimensional concept, one that can be viewed in terms of the interlocking steps or stages involved in producing a given data set. She specifies a number of characteristics that, if present, would do much to assure the collection of high-quality survey data (e.g., probability sampling; conceptual clarity; operational definitions that fit concepts; reporting by the most knowledgeable respondent; accurately coded and weighted data; small sampling variances; verification procedures that show little inconsistency (Bailar, 1984, p. 43)). Other researchers (Anderson *et al.*, 1979; Groves, 1989) discuss survey quality in terms of the various types of errors that detract from data accuracy. Groves (1989), for example, distinguishes between nonobservation error (i.e., coverage error, nonresponse error, and sampling error) and measurement error (i.e., error arising from the interviewer, respondent, questionnaire, and mode of data collection).

Although the approaches are different, Bailar's and Groves's conceptualizations of survey quality are helpful in making the concept less abstract and in setting limits for this chapter. With regard to the latter, we will focus on methods and techniques that enable researchers to detect and potentially reduce sources of measurement error. (We will not be addressing the effects of nonobservation errors on survey quality.) For example, cognitive interviews and respondent debriefing techniques (e.g., follow-up probe questions) can be used to determine when survey concepts are misunderstood. Interactional coding, in addition to indicating where interviewers have difficulty reading questions as worded or where respondents have difficulty providing adequate answers, can also be used to monitor interviewer performance and evaluate mode effects. When detected by such techniques, the presumption is that problems will have a negative effect on data quality; in some cases, howevèr, the magnitude of the effect may be difficult to assess quantitatively. Nevertheless, once identified, it is assumed that questionnaire designers (i.e., teams of subject matter specialists, behavioral scientists, and survey methodologists) can attenuate problems via question modifications (e.g., changes in wording or question structure) or other strategies (e.g., interviewer training, mode changes). The net result, if assumptions prove true, would be an increase in data quality via a reduction in measurement error. (See Section 24.3.1 for more on the topic of identifying and correcting problems with survey questions.)

24.1.2 Evaluating Survey Questionnaires, Tools of the Trade

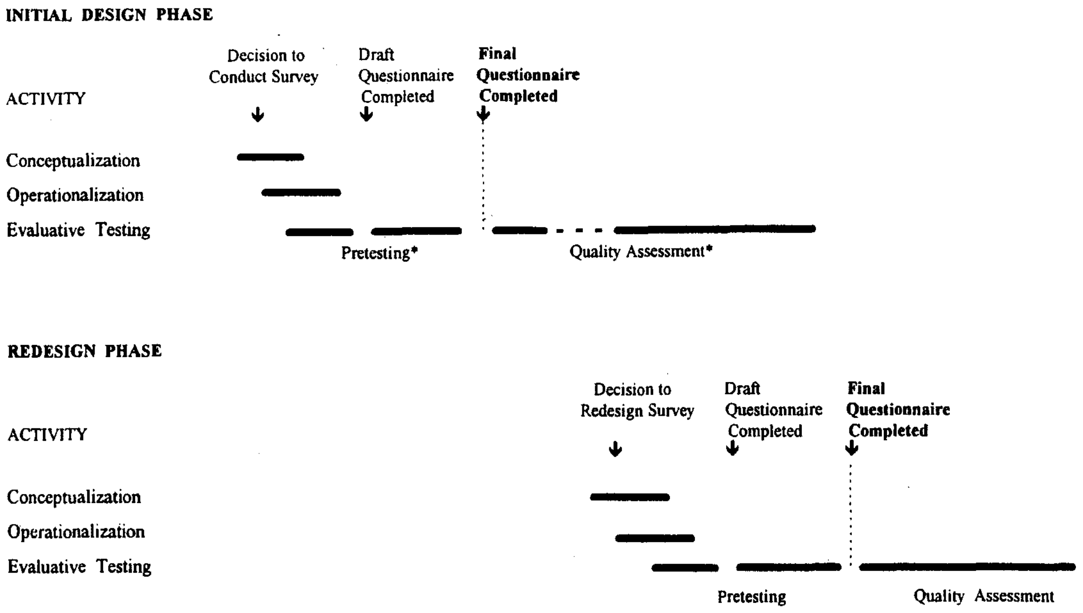
There is an expanding literature (e.g., DeMaio, *et al.*, 1993; Forsyth and Lessler, 1991; Willis, 1994) on the methods and techniques used for pretesting survey questions (see Table 24.1). These methods and techniques can be differentiated in terms of where evaluative data are usually collected (office, laboratory, or field), the purpose of the evaluation (pretesting (PT) or quality assessment (QA)), and the source (S) and target (T) of the analytical data (respondents, interviewers, others). The information in the *purpose* column highlights a point seldom made in the literature: many of the same methods used to develop and field test survey questionnaires can be used—and have been used—for the purpose of evaluating how well questionnaire items are working *after* the questionnaire has been finalized and put into production. Information contained in Table 24.1 could be used in the early stages of planning a questionnaire evaluation plan. Researchers wishing to develop a multimethod quality assessment program could use information in the *source and target* columns to identify techniques (e.g., debriefings, behavior coding, cognitive forms appraisal) that draw information from a variety of sources (i.e., interviewers, respondents, experts), and, in so doing, allow for multiple perspectives in identifying potentially problematic survey questions.

We believe that the conceptual distinction between pretesting and quality assessment research rests with the *status of the questionnaire.* If the questionnaire is in an early developmental stage or a field testing stage when evaluative research takes place, and if changes to the questionnaire can still be made after testing, then we refer to that evaluative work as *pretesting research* (see Figure 24.1). If the questionnaire is currently in use and the purpose of the evaluative research is to determine, for example, the extent to which survey questions accurately measure the concepts they are intended to measure, then we would classify that evaluative work as *quality assessment research.* (Other uses are noted in Section 24.3.) These two designations of survey research are not exhaustive. For example, not considered above is methodological research

Table 24.1 Some Methods and Techniques Used to Evaluate Survey Questions

	Location of		Source (S) and target (T) of analytical data			
Method/Technique	data collection	Purpose	Interviewers	Respondents	Other S/T	Survey Qs
Cognitive/Intensive interviews						
(1) using think-aloud technique	Lab	PT		S		T
(2) using probing technique	Lab	PT		S	T (concepts)	T
(3) using other techniques	Lab	PT		S		T
Debriefings						
(1) post-survey follow-up probes	Field	PT/QA		S	T (concepts)	T
(2) debriefing questionnaires	Field	PT/QA	S	S		T
(3) focus groups	Field/Lab	PT/QA	S	S		T
Experiments (e.g., split ballot tests)	Field/Lab	PT/QA	S	S	S (context)	T
Expert reviews						
(1) expert panels	Office	PT			S (experts)	T
(2) cognitive forms appraisals	Office	PT/QA			S (experts)	T
Interaction coding						
(1) behavior coding	Field	PT/QA	S	S	S (context)	T
(2) conversation analytic coding	Field/Lab	PT/QA	S	S	S (context)	T
(3) protocol coding	Lab	PT/QA	S	S		T
Item nonresponse analysis	Field	PT		S		T
Reinterviews	Field	QA	S	S	S (context)	T
Response distribution analysis	Field	PT		S	S (design team)	T
Vignettes	Lab/Field	PT	S	S	T (concepts)	T

Note: PT refers to pretesting and QA refers to quality assessment.
Primary Sources: DeMaio *et al.*, 1993; Forsyth and Lessler, 1991; Willis, 1994.



* **Pretesting**, as we are using the term, refers to evaluative testing conducted prior to finalizing a particular questionnaire and includes developmental testing and field testing. **Quality assessment** (e.g., reinterview) refers to evaluative testing conducted after a particular questionnaire has been finalized. Quality assessment, when done, may or may not be a continuous process.

Figure 24.1 A simplified timeline of the survey questionnaire design and redesign process.

conducted primarily to demonstrate the utility of a particular technique (e.g., behavior coding) for a specific purpose (e.g., evaluating interviewer performance; see Mathiowetz and Cannell, 1980). Moreover, these two designations are not always easy to distinguish. For example, an ongoing quality monitoring program might reveal a serious flaw with a particular survey question that requires a more intrusive intervention (e.g., a change in question wording) than normally used (e.g., targeted interviewer training). Lastly, as depicted in Figure 24.1, quality assessment research and pretesting research can take place simultaneously for the same survey (e.g., during questionnaire redesign efforts). (We refer to this figure again in Section 24.3.1.)

As Table 24.1 illustrates, survey evaluation data can be collected in various locations: laboratories, offices, and field settings. In the following sections, we will consider first laboratory- and office-based methods and research, and then field-based methods and research.

24.1.3 Laboratory- and Office-Based Methods and Research

Laboratories provide investigators with a controlled environment within which to conduct questionnaire-related evaluative research (see DeMaio and Rothgeb, 1996; DeMaio *et al.*, 1993; Dippo *et al.*, 1993; Dippo and Norwood, 1992; Sirken 1991; Willis *et al.*, 1991, for examples and reviews of research conducted at U.S. governmental research laboratories). Laboratory-based methods (e.g., cognitive interviews) generally focus on psychological aspects of the question-answering process (e.g., cognitive processes such as comprehension; motivational processes such as satisficing or selective reporting; see Forsyth and Lessler, 1991; Lessler *et al.*, 1989; Royston *et al.*, 1986; Willis, 1994). These methods have been used regularly—if not always correctly or consistently (see Blair and Presser, 1993)—by researchers to assess the understandability of survey questions and concepts. Unless there are good reasons for not doing so, researchers select and interview individuals who possess characteristics similar to those of the target population.

24.1.3.1 Cognitive and Intensive Interviews

When applied to the survey response process, the methods of cognitive and intensive interviewing rest on a multistage model of human information processing (see Willis, 1994; Willis *et al.*, 1991). This model depicts the respondent as attempting to comprehend the target question, retrieving relevant information from memory, making decisions regarding level of effort and self-presentation, and ultimately producing a response. No presumption of strict linear processing is made; for some questions, respondents presumably switch back and forth between stages. In the *cognitive interview*, the researchers attempt to identify problems with survey questions by having respondents "think out loud" as they formulate their answers (concurrent) or shortly after they formulate their answers (retrospective). Oftentimes, researchers will incorporate probing techniques into the interview by having respondents answer a limited number of general or item-specific probe questions. *Intensive interviews* assume the form of a cognitive interview, but differ from the latter in that they are supplemented with various other techniques—such as, paraphrasing, confidence ratings, response latency measures (Royston, 1989; Royston *et al.*, 1986; also see DeMaio and Rothgeb, 1996). Using such procedures, researchers gain insights into where respondents are experiencing cognitive difficulties and where there may be discrepancies between question intent (survey sponsor, question designer) and question interpretation (respondent).

24.1.3.2 Other Laboratory- and Office-Based Methods

In addition to cognitive and intensive interviews, there are a variety of other methods (e.g., expert evaluations, rating tasks) that are well suited to the research laboratory. Interested readers should refer to Forsyth and Lessler (1991) for a discussion of these methods. Office-based research methods are now appearing (e.g., conversation analytic coding; automated coding) that reflect the development of specialized analytical techniques (e.g., linguistic analysis, computerized coding). Other methods take advantage of the knowledge of experts who have proven expertise in designing questionnaires (e.g., cognitive forms appraisal, expert panels).

With the relatively recent dawn of the CASM movement and the establishment of three cognitive laboratories within the U.S. Federal statistical system (at the National Center for Health Statistics, the Bureau of Labor Statistics, and the Bureau of the Census), cognitive research in governmental surveys has proliferated. To give readers a sense of how these methods are used, we describe briefly some of the more innovative studies we have found in the literature.

24.1.3.3 Pretesting Research Examples

Blixt and Dykema (1993) have developed an innovative pretesting method, called *systematic intensive interviewing*, that integrates intensive cognitive interviews—involving think-aloud procedures, paraphrasing, memory probes, and other probing techniques—with behavior coding as a way of identifying problematic survey questions. In this study, a variety of different probes were asked during the cognitive interviews: definition probes ("When I use the word 'nutrition', what does that mean to you?"), frame-of-reference probes ("When I said 'other people,' what did that mean to you?"), and motivation-relevant probes ("Do you find it embarrassing to talk about how often you drink?"). Trained researchers conducted intensive interviews, which were audiotaped and later coded using: (1) standard respondent behavior codes (e.g., qualified response, request for clarification), and (2) specialized codes to reflect content generated in response to the probes mentioned above (e.g., critiques of the survey question and response options). Blixt and Dykema review data for two problematic questions and their analysis suggests that systematic intensive interviewing can be an effective method for identifying both cognitive and motivational problems with survey questions.

Bolton (1993) has developed a method of analyzing survey questions, called *automatic coding*, that integrates cognitive interviews and computerized content analysis. The coding scheme is well grounded in cognitive theory and matches instances of five verbal categories (i.e., repeat, forget, confidence, "can't say," and "don't know") and four nonverbal cue categories (e.g., pauses, broken utterances) with the contents of cognitive interview transcripts. She then used factor analysis (i.e., factor scores for comprehension, retrieval, judgment, and response) as a means for identifying specific cognitive problems with target questions. In this research, Bolton evaluated alternative customer satisfaction surveys and contrasted automatic coding with observational monitoring (i.e., a technique similar to behavior coding). She found that automatic coding was better at detecting comprehension, retrieval, and judgment problems, but that observational monitoring was better at detecting response difficulties.

24.1.3.4 Quality Assessment Research Examples

Forsyth *et al.* (1992) and Forsyth and Hübbard (1992) have developed a technique, called *cognitive forms appraisal*, that we believe holds great promise as a questionnaire evaluation tool—both for quality assessment and for pretesting. Items on a questionnaire are coded by experts using a coding scheme, which is grounded in a cognitive model of the survey response process. Detailed coding categories were developed to assess the demands that survey questions place on an individual's comprehension, interpretive, memory, judgment, and response gėneration processes. Forsyth *et al.* used this coding scheme to evaluate questions on the National Household Survey on Drug Abuse and found evidence of several types of problems (e.g., vague or ambiguous terminology, response categories with hidden definitions); later, the investigators were able to validate their findings using a small number of think-aloud interviews. (For another example of quality assessment research, see Dykema *et al.*, Chapter 12.)

24.1.4 Field-Based Methods and Research

Field-based methods draw information from two principal sources, interviewers and respondents (DeMaio *et al.*, 1993) and, as the name suggests, these techniques are conducted in the field, usually at or very near to where the survey is actually being conducted. We will consider four groupings of methods: interviewer debriefings, respondent debriefings, interaction analysis, and reinterview programs.

24.1.4.1 Interviewer Debriefings

The experienced field interviewer is usually one of the first persons to know how well survey questions are being understood by respondents (Converse and Schuman, 1974; DeMaio, 1983; cf. Bischoping, 1989). This expert knowledge can be tapped by a variety of methods; three of the more common techniques are to use rating forms, focus groups, and interviewer debriefing questionnaires (Fowler, 1989; Fowler and Roman, 1992; Esposito and Hess, 1992). The debriefing questionnaire has the advantage of being more representative of the larger group of interviewers; focus groups have the advantage of providing greater depth and insight into the questions that may be causing problems for respondents and interviewers. Though useful, one must recognize the subjective nature of interviewer debriefing data and take steps to obtain input from a variety of sources (e.g., interviewers, respondents, experts, coded interaction data from actual interviews) when attempting to identify and to amend problematic survey questions.

24.1.4.2 Respondent Debriefings

Perhaps the most obvious way to gather data on how well survey questions are understood is to ask the persons answering the questions—the respondents. Belson (1981) has done some very interesting research using follow-up probes to determine when concepts and terms are being understood in a way unintended by the survey sponsor or the questionnaire designer. Other researchers have used vignettes as a way of debriefing respondents (Martin, 1986; Polivka and Martin, 1992). Another approach would be to use focus groups for the same purpose (Palmisano, 1989). The commonality that unites each of these techniques is that respondents are providing information that can be used to assess their understanding (or misunderstanding) of survey questions.

24.1.4.3 Interaction Analysis

Interaction analysis involves monitoring survey behavior (most often in a natural survey context), coding behavioral exchanges between interviewers and respondents, and tabulating the frequency of a predetermined set of behavior codes for specific survey questions. Questions with relatively high frequencies of unacceptable behavior codes are flagged as problematic (for an example, see Dykema *et al.*, Chapter 12; for reviews, see Fowler and Cannell, 1996; Esposito *et al.*, 1994). The most commonly used application of interaction analysis, *behavior coding*, was developed by Cannell and his colleagues (e.g., Cannell *et al.*, 1968; Marquis, 1969; Cannel *et al.*, 1975). Morton-Williams and Sykes (e.g., Morton-Williams, 1979; Morton-Williams and Sykes, 1984; Sykes and Morton-Williams, 1987) have also made significant contributions to the literature.

24.1.4.4 Reinterview Programs

As the name of this method suggests, reinterviews involve conducting a second interview with a given unit (Biemer and Forsman, 1992; Cantwell *et al.*, 1992; Forsman and Schreiner, 1991). And insofar as there are a variety of purposes (e.g., to evaluate field work; to estimate response variance and response bias) and ways of conducting the second interview (e.g., same respondent vs most knowledgeable respondent; same interviewer vs different interviewer; same questions vs conceptually similar questions), one can think of reinterview programs as involving a family of closely related techniques. Depending on its purpose and scope, a reinterview program may require a greater investment of resources relative to other evaluation methodologies and may alienate some respondents (Blair and Sudman, 1993).

24.1.4.5 Pretesting Research Examples

Work by Fowler (1992) and his colleagues (Cannell *et al.*, 1989) provides a particularly instructive example of the use of behavior coding (and response distribution analysis) to evaluate unclear survey questions. These investigators developed a 60-item questionnaire comprising questions from a variety of health surveys. Fowler's research targeted seven of the more problematic questions identified through standard behavior coding procedures (e.g., "Do you exercise or play sports regularly?"). Each of these questions contained poorly defined terms or concepts (e.g., exercise). The questions were revised and incorporated into a new questionnaire that was subsequently readministered; in one case, one very long and difficult question was replaced with several shorter questions. Behavior coding and response distribution analyses revealed significant improvements for many of the questions. Even after the modifications, however, several of the revised questions had relatively high percentages of inadequate answers and requests for clarification. This outcome underscores two important points about "fixing" survey questions: (1) sometimes an attempt to fix one problem results in the creation of one or more other problems, and (2) some questions are very difficult to repair (e.g., complex questions), and doing so could affect respondent burden by increasing the total number of questions being asked in the survey.

Another study with interesting implications was conducted by Willis (1991) on a draft health questionnaire that involved behavior coding (done while the interview was in progress), interviewer debriefings, and observer debriefings. After completing 49 field interviews, observers and interviewers were debriefed on their perceptions regarding problems with specific survey questions. Willis found that the debriefing produced more reports of problematic questions (87) than behavior coding (62). Of the 94 questions identified as problematic, the two methods agreed 59 percent of the time. Rather than be overly concerned by the relatively low level of between-method correspondence, Willis made a very provocative observation: "Under circumstances in which the behavior coding appears to be more conservative than the debriefing, it is perhaps best to view those questions found to be problematic under both methods as *clear candidates for modification*, and to consider as additional possibilities those questions identified by the debriefing alone" (p. 12, italics added). We will expand on this insight later in the chapter (see Section 24.3.1).

24.1.5.6 Quality Assessment Research Examples

Although their research was clearly done with a different purpose in mind (i.e., to demonstrate the utility of two distinct pretesting methodologies), recent research by Oksenberg, Cannell, and their colleagues (Oksenberg *et al.*, 1991; Cannell *et al.*, 1989) provide some very useful examples of how to do excellent quality assessment research. In the work reported by Oksenberg *et al.*, the researchers started by generating a 60-item questionnaire covering a wide range of medical issues and question types. All questions were currently being used in major health surveys and presumably had been pretested in some manner. The research team administered the questionnaire by telephone to 164 respondents. Interviewer-respondent interactions for all interviews were behavior coded using standard procedures. The findings were startling: 60 percent of the questions tested had the inadequate answer code assigned 15 percent of the time or more; this code is given when the respondent's answer does not satisfy the question objective. Fifty percent of the questions had the request for clarification code assigned 10 percent of the time or more. In 104 interviews, four types of special follow-up probe questions (i.e., general, comprehension, retrieval and response category selection probes) were asked after target questions were asked. Of the different types of probes used, the comprehension probes proved to be the most useful. These probes uncovered problems with questions that some respondents originally answered with little difficulty; such problems go undetected by behavior coding. By recognizing and demonstrating the complementary aspects of question evaluation methods (in this case, behavior coding and follow-up probes), these researchers make a major contribution to the literature. An equally important contribution, strikingly demonstrated by the data presented above and in their report, was that conventional pretesting techniques do not always catch serious problems with survey questions.

24.2 MAKING TRANSITIONS FROM PRETESTING TO QUALITY ASSESSMENT: THE REDESIGN OF THE CURRENT POPULATION SURVEY (CPS) AS A CASE STUDY

In the mid-1980s, inspired by the increased application of cognitive psychology to survey measurement and advances in computer assisted interviewing, the U.S. Bureau of Labor Statistics (BLS) and the U.S. Bureau of the Census decided to redesign the Current Population Survey (CPS) for use in a user friendly, computer assisted interviewing environment. Both pretesting and quality assessment methodologies were used in the redesign (Campanelli *et al.*, 1991; Esposito *et al.*, 1991, 1992; Rothgeb *et al.*, 1991). During the period 1986–89, preliminary research was conducted to identify conceptual problems in the CPS that needed to be addressed in the redesign process (see Copeland and Rothgeb, 1990, for a brief review; also see Martin, 1987). A variety of techniques were used by BLS and Census Bureau researchers in conducting studies to identify problems with the CPS questionnaire:

- interviewer debriefings using focus groups with CPS interviewers (U.S. Bureau of Labor Statistics, 1988)
- respondent debriefings using focus groups, follow-up probe questions, and vignettes (Palmisano, 1989; Campanelli *et al.*, 1989; Campanelli *et al.*, 1991)

- categorical sorting tasks (Fracasso, 1989)
- field experiments (Westat/AIR, 1989a, 1989b).

Findings from these studies, and other sources (e.g., National Commission on Employment and Unemployment Statistics, 1979; U.S. Bureau of Labor Statistics, 1986, 1987, 1988), helped to set the stage for the redesign of the CPS.

24.2.1 The CPS Redesign: Pretesting Research (Phase One and Phase Two)

The CPS redesign involved three phases. The first two phases were designed to evaluate items on alternative versions of the CPS questionnaire.

24.2.1.1 Phase One

During the phase one field test, two alternative CPS questionnaires (versions B and C) were field tested with the existing CPS (version A) serving as the control. After this initial pretesting phase, the best questions from the alternative questionnaires were synthesized into a single alternative questionnaire (version D), which was tested in phase two. The following pretesting methodologies were used during this initial phase: (a) interviewer debriefings, (b) field-based respondent debriefings, (c) behavior coding, and (d) item-based response analysis.

Interviewer Debriefings

Phase one research utilized two aspects of interviewer debriefing: (1) the completion of a self-administered questionnaire, and (2) participation in focus group discussions with other interviewers. Although the two interviewer debriefing techniques 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 (A, B, or C) to their specific evaluations of a particular question, or series of questions. From interviewer debriefings, researchers were able to obtain interviewers' perceptions of problems with specific questionnaire items.

Respondent Debriefings

To obtain information about respondents' understanding of CPS questions, field-based respondent debriefings were conducted with household respondents after their fourth and final monthly interview. This postinterview consisted either of vignettes or a series of appropriate follow-up probe questions. For example, when given work or nonwork vignettes to evaluate (e.g., "In addition to attending her regular college classes, Bill (Jan) earned some money tending bar for a fraternity (sorority) party last week."), respondents were asked to classify the target person (Bill or Jan) as either working or not working. To assess whether a reported business in the household satisfied BLS criteria for a business, respondents had to answer "yes" to at least one of several follow-up probe questions (e.g., "Do you advertise the products or services of the business. for example, by displaying a sign, or listing the business in the phone book or newspapers?"). Respondents' eligibility for a set of follow-up questions was determined by their responses during the main interview. From respondent debriefings, researchers were able to obtain quantitative data regarding respondent comprehension of concepts and questions, and make revisions, as appropriate.

Behavior Coding

Using a specially developed form and coding procedures (based on work by Cannell *et al.*, 1975, 1989; Shepard and Vincent, 1991), interviewer–respondent interactions were monitored and coded by BLS and Census Bureau researchers. Procedures were designed to allow the coding of interviewer–respondent exchanges *during an actual interview*. Monitors noted whether interviewers read a question exactly as worded, with a slight change in wording, or with a major change in wording. Deviations in question wording were coded as major changes if they altered the meaning or intent of the question. For the respondent, researchers distinguished among the following behaviors: gives adequate answer, gives qualified answer, gives inadequate answer, asks for clarification, interrupts, does not know, or refuses to answer. Behavior coding helped to identify items that caused problems for interviewers (e.g., manifested by deviations from the exact question wording) or that caused problems for respondents (e.g., manifested by inadequate answers, requests for clarification, interruptions, etc.).

Item-Based Response Analysis

The purpose of nonresponse and response distribution analyses was to determine the extent to which differences in question wording or question sequencing produced different patterns of responses. *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. Refusal rates provided data on the sensitivity of particular questions. "Don't know" rates provided an indication of respondent task difficulty. *Response distributions* were generated and tabulated for all survey questions, with special attention being paid to the response distributions of comparable questions which differed across questionnaires.

Table 24.2 summarizes how the methods described above were used during phase one to select the CPS "work" question for the version D questionnaire tested in phase two. Other CPS questions were evaluated in a similar fashion. In reviewing this table (and, later, Table 24.3), the reader should take note of two things. First, with the exception of respondent debriefing data (i.e., the follow-up probe questions), evaluative methods provide only indirect information regarding survey data quality. By comparing evaluative data (e.g., behavior coding data) across alternative work questions, analysts infer which question wording is producing higher quality data (e.g., the question with higher

Table 24.2 Selecting the *Work* Question for Version D

Alternative Questions

Version A: Did you do any work at all LAST WEEK, not counting work around the house?

Version B: LAST WEEK, did you do any work for pay or profit?

Version C: LAST WEEK, did you do any work at all? Include work for pay or other types of compensation?

Goal: To select a *work* question for the version D questionnaire that best operationalizes the concept of work and that minimizes problems for respondents and interviewers. (Criteria for the concept of work include: work for one hour or more for pay or profit, pay-in-kind, or unpaid work in a family business or farm for 15+ hours during the reference week.)

Measurement Issues: To determine effects of question wording on respondents' interpretation of "work" as a concept and the reporting of work activities.

Results of Evaluative Methodologies and Techniques

A. Behavior Coding: Data analyses provide support for selecting version B question.

- Marginally significant difference among alternative versions of the work question with respect to the percentage of time interviewer read the question exactly as worded (A = 94.3%; B = 98.8%; C = 93.9%).
- Nonsignificant difference among alternative versions of the work question with respect to the percentage of respondents who gave an adequate answer to the question (A = 90.9%; B = 95.6%; C = 91.9%).

B. Interviewer Debriefings: Debriefings suggest that interviewers (and respondents) experience some difficulties with all three versions of the work question.

(1) Focus groups. Some interviewers report not liking the A question because it sounds demeaning to housewives and because it is confusing to some respondents (e.g., volunteer workers). The use of the term "profit" in the B question confuses some respondents—especially those who do not have a business. The use of the phrase "other types of compensation" in the C question confuses some respondents and some interviewers, too.

(2) Debriefing questionnaire (N = 68 interviewers). When asked what question was most difficult for them to ask, two interviewers selected the version A work question (too wordy or awkwardly worded), three selected the version B work question (confusing, ambiguous, difficult to understand), and three selected the version C question (same reasons as B). When asked what question appeared to be most difficult for respondents to answer, five interviewers selected the version B work question (confusing, ambiguous, difficult to understand) and three selected the version C work question (same reasons as B). When asked what terms or concepts were most commonly misunderstood by respondents, six interviewers mentioned "working for pay or other types of compensation"; four mentioned "working for pay or profit" or just "profit"; and four mentioned "work" or "work vs employed."

C. Respondent Debriefings: Data analyses provide some support for all three questions.

(1) Follow-up probe questions. All three work questions were effective at identifying employed persons. Differences in the percentage of *employed individuals missed* for all possible question pairings were not significant (A = 2.0%; B = 1.8%; C = 1.1%).

(2) Vignettes. No one of the three work questions was better at eliciting responses that match CPS definitions (i.e., no one question clearly outperformed the other two alternatives). Some evidence to suggest that version B question wording may be less inclusive than other alternatives, in that a higher percentage of respondents say "no" to all vignette scenarios. Version B question is less successful than alternatives in correctly classifying marginal work activities (e.g., work in the home), but better at correctly classifying nonwork activities (e.g., volunteer service).

D. Item-Based Response Analysis: Data analyses suggest that no one question is better or worse than the alternatives.

(1) Response-distribution analyses. All three work questions produced approximately the same percentage of individuals reported as working (A = 59.16%; B = 57.95%; C = 58.71%; differences in stated percentages for all possible question pairings are not significant).

(2) Nonresponse analyses. Very little item nonresponse across versions (A = 0.18%; B = 0.18%; C = 0.22%).

Recommendation (R) and Justification (J)

R: Adopt a slightly modified version of the version B work question for the version D questionnaire: "LAST WEEK, did you do ANY work for (either) pay (or profit)?" parentheticals to be read only if respondent answers "yes" to the prior question regarding a family business or farm (i.e., "Does anyone in this household have a business or farm?"). Interviewers to emphasize the reference period "LAST WEEK" and the word "ANY."

J: Response analyses and respondent debriefings were inconclusive; that is to say, there was little or no evidence to suggest that any one of the question alternatives was better or worse than the others. Behavior coding analyses provided support for selection of the version B work question. Interviewer debriefings indicated that all three work questions have problems. Some of the confusion regarding the word "profit" in the version B question is easily rectified by having that word only appear if someone in the household has a business or a farm.

percentages of exact question readings and higher percentages of adequate answers). And second, analytical methods do not always produce clear cut results, especially when the concept being measured has a multifaceted definition (e.g., the concept of work). We believe the current example is more representative of the types of problems that survey researchers encounter and solve on a regular basis.

24.2.1.2 Phase Two

During the phase two field test, a newly synthesized questionnaire (version D) was tested with the existing CPS again serving as control. The primary objective of phase two was to identify problem areas in question wording in order to finalize development of the revised CPS questionnaire. Evaluation of the data was used to determine if there were any "fatal flaws" in version D (e.g., gross errors in the way questions or concepts were being understood) and make necessary modifications. As a result, some of the methodologies were utilized differently. For example, in phase two, interviewer debriefings were conducted via focus groups only, and the analytical scope was narrower. The greatest concern was with survey questions that had direct effects on labor force classification (e.g., items having to do with work and layoff from work), and the protocol reflected this concern. Behavior coding was conducted to determine whether there were any problematic items on the version D questionnaire; items on the control questionnaire (version A) were not evaluated. Respondent debriefings (follow-up probes and vignettes) and response distribution analyses were similar to those used in phase one; and when these data were available, differences between version A and version D were examined. Relying on data generated by the various pretesting techniques described above, researchers found no evidence of serious flaws in the version D questionnaire (see Table 24.3). Only minor wording changes were recommended.

24.2.2 The CPS Redesign: Quality Assessment Research (Phase Three)

Beginning in July 1992, the revised CPS questionnaire was tested in a separate survey (12,000 households per month) conducted in parallel with the existing CPS (60,000 households per month) to determine the effect of content revisions and a new data collection technology on labor force estimates. The parallel survey (phase three) differed from the prior phases in several respects. For example, the phase three sample consisted of households from an address list, and interviewing was done in person during the first- and fifth-month interviews and by telephone during the subsequent months. In addition to assessing the effects of the new questionnaire and the new data collection technology on labor force estimates, interviewer debriefings, behavior coding, respondent debriefings, and item-based response analyses were again conducted during phase three to assess the quality of data produced by the revised questionnaire. Only now, given the distinction suggested in Section 24.1.2, researchers were no longer pretesting, but rather conducting a quality assessment of the changes made to the revised CPS questionnaire.

24.2.2.1 Phase Three Methods and Results

Both focus groups and a standardized interviewer debriefing questionnaire were used to debrief interviewers during this phase. The primary objective of interviewer debriefings was to identify problems with the revised questionnaire and with procedures that required additional interviewer training or enhancements to the interviewer manual. Focus groups were conducted after interviewers had at least three to four months experience using the revised questionnaire and the new data collection technology (e.g., laptop computers). Interviewers were asked to identify questions that caused the most difficulties for them in terms of getting adequate answers from respondents. In order to have some measure of the magnitude of the problems, interviewers were asked to use rating forms to estimate the percentage of the time they had difficulty getting an adequate answer from respondents for each of the problematic items identified. The most problematic questions, as identified by interviewers participating in the focus groups, were then included in a standardized debriefing questionnaire that was sent out to the entire staff of CPS interviewers. In this self-administered debriefing questionnaire, interviewers were asked to estimate what percentage of the time they had difficulty getting an adequate answer from respondents when asked particular questions.

During phase three, behavior coding was used to assess the quality of the revised questionnaire by examining interviewer–respondent interactions for various survey questions. It was assumed that a high percentage of interviewers reading questions exactly as worded (or with a slight change) and respondents providing adequate (or qualified) answers would contribute positively to data quality. With the consent of respondents, a selected group of interviewers audiotaped 364 parallel survey interviews. Due to the large number of taped interviews, four CPS experienced supervisory staff were trained to do the behavior coding. The coding staff used a standardized coding form that was adapted for office-based coding; in terms of content, this form was very similar to the ones used in earlier phases of testing.

Respondent debriefings were conducted during phase three to assess question comprehension and response formulation. As in other phases, follow-up probe questions were asked only of household members for whom the questions were applicable. Vignettes were not used during this phase. In addition to providing an indication of how respondents interpret the "work" question, some measures of response accuracy (e.g., missed employment data) were obtained from these debriefing questions (Rothgeb, 1994).

Response distribution analyses in phase three were conducted, in part, to determine if the revised questionnaire and new data collection technology produced response patterns consistent with those obtained in phase two. (These data were also important in comparing differences in labor force classifications produced by the existing (version A) and revised CPS questionnaires.) Item nonresponse data were also produced for selected items that were suspected to be vulnerable to higher than usual refusals or "don't know" responses. So that there would be enough cases to conduct analyses for infrequently asked items on remote question-asking paths, data were cumulated over a 12-month time period. Data gathered through other methods (e.g., respondent debriefing) were useful in explaining differences in response patterns when such differences were found (Rothgeb, 1994).

It is important to realize that, as with pretesting, measures of quality assessment provide both qualitative and quantitative data. Data quality acceptance thresholds can vary among questions being evaluated. Questionnaire items that are most critical to the construct of a concept will most likely require stricter criteria for indicating acceptable data quality. Also, it should be noted here that while the quality assessment methods discussed above frequently provide complementary data, they sometimes yield incongruent data. Given the unique perspectives of the various data sources (i.e., coders, interviewers, respondents, and questionnaire designers), the existence of some incongruencies should not be viewed as an indication that the methods are unsuitable for quality assessment purposes.

To illustrate the transition from pretesting to quality assessment, we will continue with the example of how the methods described above were used in the evaluation of the CPS "work" question (revised question: "LAST WEEK, did you do ANY work for (either) pay (or profit)?"). We choose the "work" question for illustrative purposes because it is one of the few CPS questions for which analytical data are available for all of the evaluative techniques described above. It is also the CPS question asked of the greatest universe of persons.

Table 24.3 displays evaluation data for the CPS "work" question from each of the three different test phases. Results from phase three *suggest* that the survey data obtained for this question was of fairly high quality. Behavior coding data indicate that interviewers read the revised "work" question exactly, or with only slight changes, nearly all the time. Respondents interviewed via CATI provided an adequate (or qualified) response 98 percent of the time, while those interviewed in person via CAPI provided an adequate (or qualified) response 93 percent of the time.

Data from the interviewer debriefings were mixed. Focus group participants reported that respondents appear to have difficulty answering the revised "work" question; in fact, of all the questions evaluated, they rated it the most difficult question for which to obtain an adequate answer. During the focus group session, these interviewers were asked: "What percentage of the time do you have difficulty getting an adequate answer from respondents when asking this question?" Mean and median ratings of 35.5 and 34 percent, respectively, were obtained. This same question appeared on the self-administered interviewer debriefing questionnaire, which was rated by 345 of 400 field interviewers (86 percent), and it produced a mean rating of 18.2 percent and median rating of 10 percent. It is interesting that the focus group participants, on average, reported they had difficulty getting an adequate answer to the work question nearly 36 percent of the time, yet behavior coding data indicated that respondents provided "adequate answers" to the question a very high percentage of the time (98 percent for CATI and 93 percent for CAPI). The discrepancy can be resolved by listening to actual CPS interviews. Some respondents answer the "work" question by providing relevant information ("Well, just my regular job.") or by asking a question in return ("Do you mean my regular job?"). (This would appear to be an example of "reporting" behavior, i.e., answering

Table 24.3 Selected Results for the "Work" Question from the Pretest (PT) and Quality Assessment (QA) Phases of the CPS Redesign

					Respondent debriefing	Behavior coding		Response distribution
Test/Design	Dates	Sample size	Questionnaire version	Interviewer debriefing	% Missed employment	INT Code (% E + mC)	RSP Code (% AA + qA)	% Yes (% NR)
Phase 1 CATI/RDD (PT)	July 1990–January 1991	70,000 hhlds Cumulative All versions	A	see Table 24.2	2.0%	94% CATI	91% CATI	59.16% (0.18%)
			B	see Table 24.2	1.8%	99% CATI	96% CATI	57.95% (0.18%)
			C	see Table 24.2	1.1%	94% CATI	92% CATI	58.71% (0.22%)
Phase 2 CATI/RDD (PT)	July 1991–October 1991	32,000 hhlds Cumulative Both versions	A	—	3.8% (2.2% paid)	—	—	57.74% (0.15%)
			D	FG: "just my job"	2.6% (2.0% paid)	100% CATI	95% CATI	57.01% (0.08%)
Phase 3 CATI/CAPI address list sample (QA)	July 1992–December 1993	144,000 hhlds Cumulative (1993 only)	Revised questionnaire	FG: 35.5% IDQ: 18.2%	2.9% (1.6% paid)	100% CATI 99% CAPI	98% CATI 93% CAPI	58.58% (0.16%)

Abbreviations: FG refers to focus-group data; IDQ refers to interviewer-debriefing-questionnaire data; INT refers to interviewer and RSP to respondent; (% E + mC) refers to the percentage of exact and minor-change question readings; (% AA + qA) refers to the percentage of adequate and qualified answers; (% NR) refers to the nonresponse percentage (i.e., refusals and "don't know" responses); hhlds indicates households.

a question by providing relevant information rather than a direct answer; see Schaeffer *et al.*, 1996, for a discussion.) When they hear such responses, most interviewers either probe or check the "yes" response option. The behavior coders, based on our training, coded such responses as adequate answers, because the implication of either response is that the respondent had a job and most probably worked at that job last week. In retrospect, however, such responses (as interviewers note) should perhaps be considered as only marginally adequate. In its present context, the question does appear to confuse some respondents; but again, other data suggest that we are obtaining accurate information from this question.

The findings discussed above provide a good example of how important it is to utilize more than one evaluation method. While interviewers are an invaluable source of information regarding problems they or respondents are having with survey questions, sometimes the information they provide is colored by the most salient interview situations, which may not always be typical. If assessment of data quality for the "work" question had been based solely on the results of interviewer focus groups, the sponsor (BLS) may have been alarmed by the results. Fortunately, more objective quality assessment data were available from behavior coding and other evaluation methodologies, such as respondent debriefing and response distributions. Differences between the mean and medians obtained from the interviewer focus groups versus those from the standardized self-administered questionnaire suggest that collective behavior may influence the reporting of problems. Some focus group participants may not have originally thought the "work" question was problematic; hearing others report it as problematic, however, may have influenced their ratings of the item. It is also possible that some interviewers may have forgotten having difficulty with this question, recalling their experiences only after being cued by others who identified the question as problematic during group discussion. In contrast, the debriefing questionnaires were completed independently; in the absence of cues provided by others, interviewers may have underestimated the frequency of problems experienced with this question. We suspect the true level of difficulty with the "work" question probably lies somewhere in between the medians obtained from the debriefing questionnaire (10 percent) and the focus groups (34 percent).

Results from respondent debriefing analyses suggest that the revised "work" question is producing high-quality data. Through the years, there had been concern that the classification of marginal work activities, unpaid work in a family business or farm, and work in the underground economy may have been underreported in the CPS (Martin and Polivka, 1992). The revised questionnaire was designed to improve reporting of these types of work. To obtain estimates of missed employment, persons not reported as employed in the main survey were asked in the respondent debriefing if they had done any work at all during the reference week, even for only a few hours. This was intended to obtain estimates of persons that may not have been reported as working when questions in the main survey were asked, but who had actually performed some type of work activity during the reference week. Only 1.6 percent of persons. not reported as employed in the main survey, were reported in the respondent debriefing as having done some paid work during the reference week. (Including unpaid and paid work, 2.9 percent were reported as working.) These data indicate that missed employment is somewhere between 1.6 and 2.9 percent. It should be noted that reported missed employment was slightly higher during phase three than during pretesting (phases one and two). However, the samples for phases one and two did not include nontelephone households, whereas the sample for phase three did. It is likely that nontelephone households would have a higher proportion of persons engaged in casual, informal, or marginal work activities than telephone households; such activities may not be reported as work in the main survey and would consequently result in a higher number of reports of missed employment.

Response distribution data revealed that 58.6 percent of persons for whom this initial work question was asked were reported as working last week. These data were consistent with estimates obtained during previous phases. Item nonresponse was virtually nonexistent for this question.

Taken as a whole, analytical data collected in phase three suggest that: (a) most of the items on the revised questionnaire were being read by interviewers as worded, (b) most of the items were being understood by respondents as intended, and (c) labor force misclassification (e.g., as measured by percentage of missed employment and employment to population ratios) was relatively low for the revised questionnaire (Rothgeb, 1994; Polivka, 1994; Dippo *et al.*, 1995). We should note that other more standard measures were also used to monitor the data collection process (e.g., gross and net difference rates from reinterview; noninterview rates by questionnaire, item, and interviewer).

24.3 CONCLUDING REMARKS

In the two preceding sections, we have provided some background material on various methods that are available for evaluating survey questionnaires (i.e., via pretesting and quality assessment research). In addition, we have tried to make that material more palpable by providing a case study of how various methodologies were actually used in evaluating the redesigned CPS questionnaire. In this final section, we delineate the broad outline of an idealized quality assessment program for major social and economic surveys, address several pragmatic issues associated with such a program, and consider some of the costs and benefits associated with quality assessment research.

Before proceeding, however, we would like to mention several uses to which a quality assessment program might be put. One use might be to suggest ideas for enhancing interviewer training. A second use might be to check the reliability of pretest results using the full production sample (e.g., phase three of the CPS redesign). A third use might be to assess the degree to which *any existing* *questionnaire* is providing quality data; if serious data quality problems are uncovered, survey sponsors may choose: (1) to initiate a formal plan for pretesting a redesigned questionnaire, or (2) to make modifications based on quality assessment data only.

24.3.1 An Idealized Quality Assessment Program

As a means of contributing to the validity and reliability of questionnaire data, we advocate a comprehensive, ongoing, multimethodology, multitechnique quality assessment program for surveys that provide key social and economic indicators. By *comprehensive*, we mean a research program that investigates the various and interrelated sources of measurement error (i.e., interviewers, respondents, data collection contexts; see Esposito and Jobe, 1991), and one that incorporates laboratory, office, and field research. By *ongoing*, we mean a program that incorporates regular or periodic evaluations of the survey instrument; this element is consistent with the idea of continuous measurement as described by Dippo (Chapter 20). By *multimethodology*, we mean a program that utilizes analytical data from multiple methods and multiple sources: (1) interviewers, (2) respondents, (3) survey sponsors and designers, and (4) natural context interactions between interviewers and respondents. And by *multitechnique*, we mean a program that incorporates multiple techniques (e.g., focus groups, debriefing questionnaires), when resources are available, for collecting information from the sources noted above. In our view, the multimethodology criterion is critical, because it provides researchers with evaluative data from four key perspectives; the chances of making poor decisions regarding survey questions are minimized when all perspectives are considered, not just one or two. The multitechnique criterion is also important, because it enables researchers to build on the strengths of individual techniques while, at the same time, compensating for the unique weaknesses associated with each.

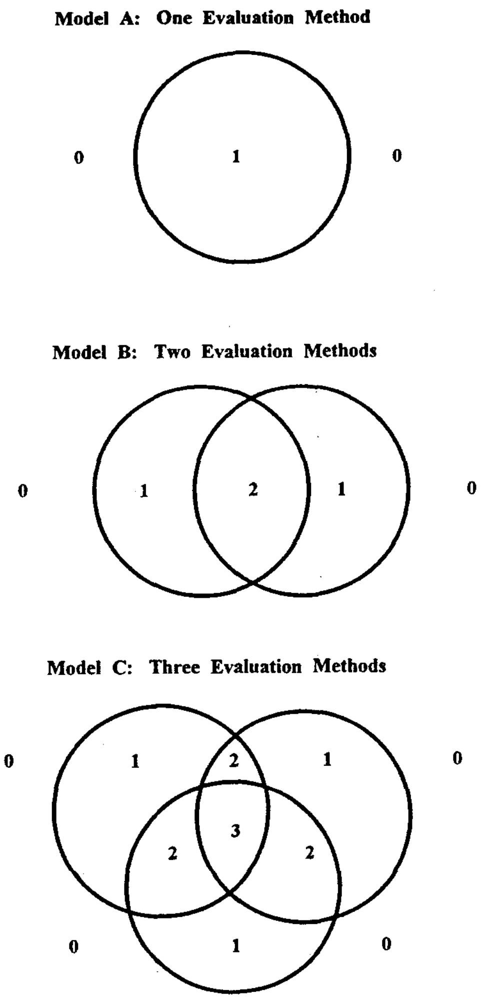
It is recognized that the ongoing program of research activities outlined above could be costly; but in considering the costs of producing quality data, one must also consider the costs (e.g., to users) of *not* producing quality data. Even so, we recognize that any quality assessment program, given the current fiscal environment, must be reasonably economical to implement. With that constraint in mind, we propose a flexible program that integrates laboratory
* and ~~behavior coding~~ in a symbiotic way. The core of the program would consist of three field methodologies (interviewer debriefings, respondent debriefings, and behavior coding) and various laboratory- and office-based techniques; such a program would be designed to supplement rather than replace existing quality assessment measures (e.g., a reinterview program). The field methodologies used in a quality assessment program should: (a) inform researchers as to where questionnaire-related problems exist, (b) provide explanations for those problems, and (c) suggest hypotheses regarding the causes of problems that can be tested under controlled laboratory conditions. In addition to testing the causal hypotheses, laboratory- and office-based research (e.g., cognitive interviews, cognitive forms appraisals) is likely to generate its own set of hypotheses which could be addressed in subsequent field research, if resources permit.

Erratum:
"... that integrates laboratory and field research in a symbiotic way."

These recommendations in hand, we now wish to raise and address three pragmatic issues regarding the quality assessment program outlined above:

1. *Given a specific evaluation method, how does a researcher determine when a "problem" exists with a given questionnaire item?* In practice, decisions as to what constitutes a "problem" with a particular questionnaire item really start with the survey concepts and the way those concepts have been operationalized in a given question (see Figure 24.1). If substantive concepts are poorly defined or undefined, the researcher has no way of judging how well specific concepts have been operationalized in a given question and no means of accurately determining if respondents understand the question. If there are no explicit question objectives, then researchers have no basis on which to evaluate a respondent's answer. (See Schwarz (Chapter 1) and Hox (Chapter 2) for an extended discussion of these issues.) Even with clear question objectives, however, establishing criteria for what constitutes a problem with a given question tends to be a subjective process; moreover, these criteria often vary across researchers and within methods. In the case of behavior coding, for example, criteria for identifying problematic questions are suggested by researchers who have substantial experience with the technique. Cannell *et al.* (1989) have flagged questions as problematic if specific behavior codes (e.g., requests for clarification) appear 15 percent or more of the time the question is asked. Many practitioners follow this rule of thumb; however, other researchers (e.g., Esposito *et al.*, 1991) have used more stringent criteria (e.g., flagging a question as problematic if adequate answers are not obtained from respondents at least 90 percent of the time the question is asked). The analysis of data from follow-up probe questions is somewhat less subjective in that probe questions are typically designed as alternative measures of the concept under consideration, and any substantial discrepancy in response distributions between the probe question and the target question is generally taken as a sign of trouble. With other techniques, interviewers or respondents identify problems by recalling the difficulties they experienced with a given questionnaire item, and then researchers must decide how much weight to assign to such judgments. In sum, based on experience (Esposito *et al.*, 1991, 1992) and our reading of the literature (e.g., Cannell *et al.*, 1989; Presser and Blair, 1994; Willis, 1991), we believe all evaluative methods have inherent weaknesses; moreover, deciding what is or is not a problematic question tends to be a subjective process. This is why we regard single method evaluation plans to be very risky.

2. *Given the use of multiple evaluation methods (e.g., behavior coding, follow-up probe questions, interviewer debriefing questionnaire), how does the researcher identify "problematic" questionnaire items?* One strategy that might be useful—at least until other strategies are developed—is to rely on what we refer to as the *relative confidence model*. This model is an extension of an idea, suggested



Model A: One Evaluation Method

Model B: Two Evaluation Methods

Model C: Three Evaluation Methods

Figure 24.2 A relative confidence model for identifying problematic survey questions.

by Willis (1991), which we interpret as follows: different evaluative methods can be viewed as complementing one another; and when used in combination, multiple methods may provide a more accurate overall means of identifying problematic questions than single methods alone. We use Venn diagrams to illustrate the model (see Figure 24.2). Each circle represents a different questionnaire evaluation method drawing information from a *different source* (e.g., interviewers, respondents, experts). In Model A, a single evaluation method has identified a certain *group of questions* as problematic (area 1). Given this single method, we have no basis for viewing the questions outside the circle as problematic (area 0). In Model B, two evaluation methods are used, and three areas circumscribed. The questions falling in area 2 have been identified as problematic by both methods; the questions falling in area 1 have been identified as problematic by one method only; and the questions falling outside these areas have not been identified as problematic by either method (area 0). *In selecting questions for review and possible revision*, we would feel most confident selecting area 2 questions as problematic, and somewhat less confident in selecting area 1 questions. Model C (three evaluation methods) follows the same logic. Our confidence in correctly selecting problematic questions for review and revision would be greatest for area 3 questions, and would decrease incrementally for areas 2 and 1, respectively. (Please note that the logic supporting the relative confidence model does not generalize as well to multitechnique comparisons that draw evaluative information from a single source (e.g., using debriefing questionnaires and focus groups to gather information from interviewers only). We would expect greater overlap between circles, but our analysis would be limited to the perspectives of a single source (i.e., interviewers) and to the particular weaknesses of the techniques used.)

3. *Given accurate identification of problematic questions (and the desire to make changes in the questionnaire), how does the researcher go about revising such questions?* One strategy would be to adopt a test-fix-retest model, similar to the process described above for phase one of the CPS redesign. Here one revises a question based on the problems identified during the initial pretesting stage and then retests the question in a second evaluative stage. But, occasionally, in revising a question to solve one problem, the question designer creates another. We really need a set of guidelines, based on empirical data, for designing good survey questions. Belson (1981, p. 389) offers some useful suggestions in this regard. For example, he suggests that we *avoid*: "loading up the question with a lot of different or defining terms; . . . ; the use of words that are not the usual working tools of the respondent; . . . ; giving the respondent a difficult task to perform; giving the respondent a task that calls for a major memory effort; . . . ." Another possibility would be to evaluate revised questions using cognitive forms appraisals before adding them to the host questionnaire (Forsyth *et al.*, 1992; Forsyth and Hubbard, 1992).

24.3.2 The Costs and Benefits of Quality Assessment Research

Survey quality cannot be achieved without incurring costs (Groves, 1989). The kind of quality assessment program described above for major social and economic surveys involves an ongoing commitment of staff and fiscal resources. Moreover, improvements in survey quality generally entail some change in the data collection process (e.g., question or procedural modifications). Program managers, though they welcome measures to improve quality, must also be concerned with the utility of their product to data users (e.g., time series integrity). These competing demands—to maintain quality levels and to assure the comparability of statistics across time—put program managers in a difficult position. To change or not to change, that is their dilemma, and it is not one that survey researchers should take lightly. Whenever quality-related improvements are recommended, program managers have every right to expect that survey researchers provide data on the anticipated consequences of those changes on their statistical products (e.g., the unemployment rate, consumer price index, the poverty index) and on operational aspects of the survey (e.g., data processing edits).

Of course, there are also substantial benefits to investing resources in quality assessment research. The most obvious benefits to users are more valid and reliable statistics (i.e., more accurate data) and greater confidence in the decisions that are made on the basis of those statistics. An obvious benefit for the organizations that produce these social and economic data is enhanced credibility. But there are benefits for the research community as well. There is an axiom in clinical psychology that goes something like this: if you want a clear understanding of normal human behavior, study maladaptive behavior patterns and attempt to determine their origins. A similar axiom might apply to survey response models and questionnaire design: if you want to understand how respondents answer survey questions and how to design questionnaires that produce high-quality data, identify problematic questions and attempt to determine why those problems exist. We believe that one of the main benefits of quality assessment research will be to help survey researchers to understand the processes (social and psychological) involved in responding to survey questions and, in so doing, help us all to design better questionnaires.

ACKNOWLEDGMENTS

Parts of this chapter (i.e., those pertaining to the redesign of the CPS) report on methods and research undertaken by staff at the U.S. Bureau of the Census and the U.S. Bureau of Labor Statistics (BLS). We wish to acknowledge the work of the individuals at both agencies who made the redesign a success, especially those whose work we have drawn upon and cited herein. The views expressed are attributable to the authors and do not necessarily reflect the views of the Census Bureau or the BLS. We also wish to thank Cathryn Dippo and Elizabeth Martin for very helpful comments to earlier drafts of this chapter.

REFERENCES

Anderson, R., Kaspar, J., Frankel, M.R., and Associates (1979), *Total Survey Error*, San Francisco: Jossey-Bass.

Bailar, B. (1984), "The Quality of Survey Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 43–52.

Belson, W.R. (1981), *The Design and Understanding of Survey Questions*, Aldershot, U.K.: Gower.

Biemer, P.P., and Forsman, G. (1992), "On the Quality of Reinterview Data with Application to the Current Population Survey," *Journal of the American Statistical Association*, 87, pp. 915–923.

Biemer, P.B., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds.) (1991), *Measurement Errors in Surveys*, New York: Wiley.

Bischoping, K. (1989), "An Evaluation of Interviewer Debriefing in Survey Pretests," in C. Cannell *et al.* (eds.), *New Techniques for Pretesting Survey Questions*, [Final Report], Ann Arbor, MI: Survey Research Center, University of Michigan.

Blair, J., and Presser, S. (1993), "Survey Procedures for Conducting Cognitive Interviews: A Review of Theory and Practice," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 370–375.

Blair, J., and Sudman, S. (1993), "Respondent Perceptions of Reinterview," *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC. pp. 701–716.

Blixt, S., and Dykema, J. (1993), "Before the Pretest: Question Development Strategies." *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 1142–1147.

Bolton, R., (1993), "Pretesting Questionnaires: Content Analysis of Respondents' Concurrent Protocols," *Marketing Science*, 12, pp. 280–303.

Campanelli, P.C., Martin, E.A., and Creighton, K.P. (1989), "Respondents' Understanding of Labor Force Concepts: Insights from Debriefing Studies," *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC. pp. 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, pp. 253–264.

Campanelli, P.C., Rothgeb, J.M., Esposito, J.L., and Polivka, A.E. (1991), "Methodologies for Evaluating Survey Questions: An Illustration from a CPS CATI/RDD Test." paper presented at the Annual Meeting at the American Association for Public Opinion Research, Phoenix, AZ.

Cannell, C.F., Fowler, F.J., and Marquis, K. (1968), *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.

Cannell, C., 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.F., Miller, P.V., and Oksenberg, L. (1981), "Research on Interviewing Techniques," in S. Leinhardt (ed.), *Sociological Methodology*, San Francisco: Jossey-Bass, pp. 389–437.

Cannell, C., and Oksenberg, L. (1988), "Observation of Behavior in Telephone Interviews", in R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, II, and J. Waksberg (eds.), *Telephone Survey Methodology*, New York: Wiley, pp. 475–495.

Cannell, C., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F.J. (eds.) (1989), *New Techniques for Pretesting Survey Questions*, [Final Report], Ann Arbor, MI: Survey Research Center, University of Michigan.

Cantwell, P.J., Bushery, J.M., and Biemer, P. (1992), "Toward a Quality Improvement System for Field Interviewing: Putting Content Reinterview into Perspective," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 74–83.

Converse, J.M., and Schuman, H. (1974), *Conversations at Random*, New York: Wiley.

Copeland, K., and Rothgeb, J.M. (1990), "Testing Alternative Questionnaires for the Current Population Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 63–71.

DeMaio, T.J. (1983), "Learning from Interviewers," in T.J. DeMaio (ed.), *Approaches to Developing Questionnaires*, Statistical Policy Working Paper 10, Washington, DC: Office of Management and Budget, pp. 119–136.

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

DeMaio, T., and Rothgeb, J. (1996), "Cognitive Interviewing Techniques: In the Lab and in the Field," in N. Schwarz, and S. Sudman (eds.), *Determining Processes Used to Answer Questions*, San Francisco: Jossey-Bass.

Dippo, C., Polivka, A., Creighton, K., Kostanich, D., and Rothgeb, J. (1995), "Redesigning a Questionnaire for Computer-Assisted Data Collection: The Current Population Survey Experience," unpublished report, Washington, DC: U.S. Bureau of Labor Statistics.

Dippo, C.S., and Norwood, J.L. (1992), "A Review of Research at the Bureau of Labor Statistics," in J.M. Tanur (ed.), *Questions about Questions*, New York: Russell Sage Foundation, pp. 271–290.

Dippo, C.S., Tucker, C., and Valliant, R. (1993), "Survey Methods Research at the U.S. Bureau of Labor Statistics," *Journal of Official Statistics*, 9, pp. 121–135.

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

Esposito, J.L., and Hess, J. (1992), "The Use of Interviewer Debriefings to Identify Problematic Questions on Alternate Questionnaires," paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Esposito, J.L., and Jobe, J.B. (1991), "A General Model of the Survey Interaction Process," *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, pp. 537–560.

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

Esposito, J.L., Rothgeb, J.M., Polivka, A.E., Hess, J., and Campanelli, P. (1992), "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.

Forsman, G., and Schreiner, I. (1991), "The Design and Analysis of Reinterview," in P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, pp. 279–301.

Forsyth, B.H., and Hubbard, M.L. (1992), "A Method for Identifying Cognitive Properties of Survey Items," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 470–475.

Forsyth, B.H., and Lessler, J.T. (1991), "Cognitive Laboratory Methods: A Taxonomy," in P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, pp. 393–418.

Forsyth, B.H., Lessler, J.T., and Hubbard, M.L. (1992), "Cognitive Evaluation of the Questionnaire," in C.F. Turner, J.T. Lessler, and J.C. Gfroerer (eds.), *Survey Measurement of Drug Use: Methodological Studies*, U.S. Department of Health and Human Services, Washington, DC, pp. 13–52.

Fowler, F.J. (1989), "Evaluation of Special Training and Debriefing Procedures for Pretest Interviews," in C. Cannell *et al.* (eds.), *New Techniques for Pretesting Survey Questions*, [Final Report], Ann Arbor, MI: Survey Research Center, University of Michigan.

Fowler, F.J. (1992), "How Unclear Terms Affect Survey Data," *Public Opinion Quarterly*, 56, pp. 218–231.

Fowler, F.J., and Cannell, C.F. (1996), "Using Behavior Coding to Identify Cognitive Problems with Survey Questions," in N. Schwarz, and S. Sudman (eds.), *Methods of Determining Cognitive Processes in Answering Questions*, San Francisco: Jossey-Bass.

Fowler, F.J., and Roman, A.M. (1992), "A Study of Approaches to Survey Question Evaluation," working paper for the U.S. Bureau of the Census, Washington, DC.

Fracasso, M.P. (1989), "Categorization of Responses to the Open-Ended Labor Force Questions in the Current Population Survey (CPS)," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 481–485.

Groves, R.M. (1987), "Research on Survey Data Quality," *Public Opinion Quarterly*, 51, pp. S156–S172.

Groves, R.M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, II, W.L., and Waksberg, J. (eds.) (1988), *Telephone Survey Methodology*, New York: Wiley.

Jabine, T.B., Straf, M., Tanur, J., and Tourangeau, R. (1984), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC: National Academy Press.

Jobe, J., and Mingay, D. (1991), "Cognition and Survey Measurement: History and Overview," *Applied Cognitive Psychology*, 5, pp. 175–192.

Lessler, J., Tourangeau, R., and Salter, W. (1989), *Questionnaire Design in the Cognitive Research Laboratory: Results of an Experimental Prototype*, Vital and Health Statistics, Series 6, Number 1, DHHS Publication No. (PHS) 89-1076, Washington, DC: Government Printing Office.

Marquis, K. (1969), "Interviewer-Respondent Interaction in a Household Interview," *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 24–30.

Martin, E. (1986), "Report on the Development of Alternative Screening Procedures for the National Crime Survey," unpublished report, Washington, DC: Bureau of Social Science Research.

Martin, E. (1987), "Some Conceptual Problems in the Current Population Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 420–424.

Martin, E., and Polivka, A.E. (1992), "The Effect of Questionnaire Redesign on Conceptual Problems in the Current Population Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 655–660.

Mathiowetz, N.A., and Cannell, C.F. (1980), "Coding Interviewer Behavior as a Method of Evaluating Performance," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 525–528.

Morton-Williams, J. (1979), "The Use of 'Verbal Interaction Coding' for Evaluating a Questionnaire," *Quality and Quantity*, 13, pp. 59–75.

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

National Commission on Employment and Unemployment Statistics (1979), *Counting the Labor Force*, Washington, DC: Government Printing Office.

Oksenberg, L., Cannell, C., and Kalton, G. (1991), "New Strategies for Pretesting Questionnaires," *Journal of Official Statistics*, 7, pp. 349–365.

Palmisano, M. (1989), "Respondent Understanding of Key Labor Force Concepts Used in the CPS," unpublished paper, Washington, DC: U.S. Bureau of Labor Statistics.

Polivka, A.E. (1994), *Comparison of Labor Force Estimates from the Parallel Survey and the CPS During* 1993: *Major Labor Force Estimates*, CPS Overlap Analysis Team Technical Report 1, Washington, DC: U.S. Bureau of Labor Statistics and U.S. Bureau of the Census.

Polivka, A.E., and Martin, E.A. (1992), "The Use of Vignettes in Pretesting and Selecting Questions," paper presented at the annual meeting of the American Association of Public Opinion Research, St. Petersburg, FL.

Presser, S., and Blair, J. (1994), "Survey Pretesting: Do Different Methods Produce Different Results?," in P.V. Marsden (ed.), *Sociological Methodology*, Volume 24, Oxford: Basil Blackwell, pp. 73–104.

Rothgeb, J.M. (1994), *Revisions to the CPS Questionnaire*: *Effects on Data Quality*, CPS Overlap Analysis Team Technical Report 2, Washington, DC: U.S. Bureau of the Census and the U.S. Bureau of Labor Statistics.

Rothgeb, J.M., Polivka, A.E., Creighton, K.P., and Cohany, S.R. (1991), "Development of the Proposed Revised Current Population Survey Questionnaire," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 56–65.

Royston, P. (1989), "Using Intensive Interviews to Evaluate Questions," *Conference Proceedings*: *Health Survey Research Methods*, DHHS Publication No. (PHS) 89-3447, Washington, DC: National Center for Health Services Research and Health Care Technology Assessment.

Royston, P., Bercini, D., Sirken, M., and Mingay, D. (1986), "Questionnaire Design Research Laboratory," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 703–707.

Schaeffer, N.C., Maynard, D.W., and Cradock, R.M. (1996), "From Paradigm to Prototype and Back Again: Interactive Aspects of 'Cognitive Processing' in Standardized Survey Interviews," in N. Schwarz, and S. Sudman (eds.), *Determining Processes Used to Answer Questions*, San Francisco: Jossey-Bass.

Shepard, J., and Vincent, C. (1991), "Interviewer–Respondent Interactions in CATI Interviews," *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, pp. 523–536.

Sirken, M. (1991), "The Role of a Cognitive Laboratory in a Statistical Agency," seminar on Quality of Federal Data (Part 2 of 3), Statistical Working Paper 20, Washington, DC: Office of Management and Budget, pp. 268–277.

Sykes, W., and Morton-Williams, J. (1987), "Evaluating Survey Questions," *Journal of Official Statistics*, 3, pp. 191–207.

Tourangeau, R. (1984), "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, pp. 73–100.

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

U.S. Bureau of Labor Statistics (1986), "Report of the BLS-Census Bureau Questionnaire Design Task Force," staff report, Washington DC: U.S. Department of Labor, Bureau of Labor Statistics.

U.S. Bureau of Labor Statistics (1987), "Second Report of the BLS-Census Bureau Questionnaire Design Task Force," staff report, Washington DC: U.S. Department of Labor, Bureau of Labor Statistics.

U.S. Bureau of Labor Statistics (1988), "Response Errors on Labor Force Questions Based on Consultations with Current Population Survey Interviewers in the United States," paper prepared for the OECD Working Party on Employment and Unemployment Statistics, Washington, DC: U.S. Bureau of Labor Statistics.

Westat/AIR (1989a), *Research on Hours of Work Questions in the Current Population Survey*, Final Report, Bureau of Labor Statistics Contract No. J-9-J-8-0083, Rockville, MD: Westat, Inc.

Westat/AIR (1989b), *Research on Industry and Occupation Questions in the Current Population Survey*, Final Report, Bureau of Labor Statistics Contract No. J-9-J-8-0083, Rockville, MD: Westat, Inc.

Willis, G. (1991), "The Use of Behavior Coding to Evaluate a Draft Health-Survey Questionnaire," paper presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Willis, G. (1994), *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.

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, pp. 175–192.

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Survey Measurement and Process Quality

Edited by

LARS LYBERG
PAUL BIEMER
MARTIN COLLINS
EDITH DE LEEUW
CATHRYN DIPPO
NORBERT SCHWARZ
DENNIS TREWIN

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper.

Library of Congress Cataloging in Publication Data:

Survey measurement and process quality / Lars Lyberg . . . [et al.]
p. cm.—(Wiley series in probability and statistics.
Applied probability and statistics)
Includes index.
ISBN 0-471-16559-X (cloth : alk. paper)
1. Social sciences—Statistical methods. 2. Surveys—Methodology.
3. Social sciences—Research—Evaluation. 4. Surveys—Evaluation.
I. Lyberg, Lars. II. Series.
HA29.S843 1996
300′.723—dc20 96-44720
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

INTRODUCTION

SECTION A. QUESTIONNAIRE DESIGN

SECTION B. DATA COLLECTION

SECTION C. POST SURVEY PROCESSING AND OPERATIONS

SECTION D. QUALITY ASSESSMENT AND CONTROL

SECTION E. ERROR EFFECTS ON ESTIMATION, ANALYSES, AND INTERPRETATION

Contributors

Reginald P. Baker, National Opinion Research Center, Chicago, Illinois, U.S.A.

Alison K. Baldwin, National Opinion Research Center, Chicago, Illinois, U.S.A.

Francesca Bassi, University of Padua, Padua, Italy

Mary K. Batcher, Internal Revenue Service, Washington, DC, U.S.A.

Jelke G. Bethlehem, Statistics Netherlands, Voorburg, The Netherlands

Paul P. Biemer, Research Triangle Institute, Research Triangle Park, North Carolina, U.S.A.

Steven Blixt, MBNA-America, Wilmington, Delaware, U.S.A.

Bill Blyth, Taylor Nelson AGB, London, United Kingdom

Pamela Campanelli, Social Community Planning Research, London, United Kingdom

Noel Chavez, University of Illinois, Chicago, Illinois, U.S.A.

Michael Colledge, Australian Bureau of Statistics, Belconnen, Australia

Martin Collins, City University Business School, London, United Kingdom

Frederick Conrad, Bureau of Labor Statistics, Washington, DC, U.S.A.

Mick P. Couper, University of Michigan, Ann Arbor, Michigan, U.S.A.

Edith de Leeuw, Vrije Universiteit, Amsterdam, The Netherlands

Don A. Dillman, Washington State University, Pullman, Washington, U.S.A.

Cathryn S. Dippo, Bureau of Labor Statistics, Washington, DC, U.S.A.

Jennifer Dykema, University of Wisconsin, Madison, Wisconsin, U.S.A.

James L. Esposito, Bureau of Labor Statistics, Washington, DC, U.S.A.

Luigi Fabbris, University of Padua, Padua, Italy