Journal of Official Statistics, Vol. 20, No. 2, 2004, pp. 143–183

Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study

James L. Esposito[1]

This article summarizes a series of three biennial evaluations of a labor force questionnaire that collects data on worker displacement. Adopting a dichotomy for evaluation research that draws a distinction between questionnaire *pretesting* (developmental/pre-implementation evaluations) and *quality assessment* (post-implementation evaluations), the first two evaluations in the series represent quality assessment research. The third evaluation is somewhat unusual in that it can be classified as *both* pretesting and quality assessment research. Though the scope of work for each evaluation differed, three standard methods for evaluating questionnaires were used during *each* of these efforts: interviewer debriefings, interaction/behavior coding, and respondent debriefing. It should be noted that this series of studies was not iterative by design – it evolved as such due to unforeseen circumstances and, in the process, yielded unanticipated benefits.

Key words: Behavior coding; displaced workers; focus groups; measurement error; pretesting; respondent debriefing.

1. Introduction and Objectives

Studies that describe iterative, multiple-method questionnaire evaluation research (e.g., Schaeffer and Dykema 2004) are fairly rare in the survey methodology literature. Such research, however, holds great promise for understanding the strengths and weaknesses of various evaluation methods and for assessing the conceptual foundations of the target survey. This article summarizes a series of three biennial evaluations of the Displaced Worker Supplement (DWS), a governmental survey that collects data on worker displacement (see Section 3). It is hoped that the article contributes to questionnaire evaluation practice and theory in the following ways: (1) by documenting the benefits of iterative questionnaire evaluation research; (2) by demonstrating the utility of a multiple-method approach to evaluating questionnaires; (3) by drawing attention to the importance of clear and well-grounded conceptual specifications in minimizing measurement error;

[1] U.S. Bureau of Labor Statistics, Postal Square Building, Room 4985, 2 Massachusetts Avenue, N.E., Washington, DC, 20212, U.S.A. The views expressed in this article are those of the author and do not reflect the policies of the U.S. Bureau of Labor Statistics. This article draws heavily on two prior conference papers (Esposito 2002 and 2003). This multiphase research effort reflects the Bureau's commitment to survey evaluation research as a means towards the goal of collecting accurate and reliable labor force statistics (Abraham 1996) and it is consistent with the pretesting policy of the U.S. Bureau of the Census (1998).

Acknowledgments: This research could not have been accomplished without the contributions and insights of content specialists at the U.S. Bureau of Labor Statistics (Thomas Nardone, Francis Horvath, Steven Hipple, Jay Meisenheimer) and without the hard work and cooperation of the U.S. Census Bureau's operations and field staff. Thanks also to Nora Cate Schaeffer and Ed Robison for providing helpful comments on an earlier draft of this article.

and (4) by providing a broad organizational framework with which to address and solve problems of both a theoretical and an applied nature.

In pursuit of these objectives, an organizational framework will be presented in the next section that interrelates various phases of the questionnaire design-and-evaluation process with elements of a widely cited model of measurement error. The framework provides the structure within which one can chart the developmental history of a particular survey questionnaire and assess its potential strengths and weaknesses.

2. The Framework

The first dimension of the framework consists of a rudimentary process model that describes in very general terms how questionnaires are developed and evaluated (Esposito and Rothgeb 1997; Esposito 2002 and 2003). (For more thorough discussions of these topics, the reader has many excellent choices: Akkerboom and Dehue 1997; Converse and Presser 1986; DeMaio, Mathiowetz, Rothgeb, Beach, and Durant 1993; Forsyth and Lessler 1991; Fowler 1995; Goldenberg et al. 2002 (for an establishment survey perspective); Oksenberg, Cannell, and Kalton 1991; Platek 1985; Snijkers 2002; Sudman and Bradburn 1982; Turner and Martin 1984; and Willis, Royston, and Bercini 1991.) The model comprises eight partially recursive and overlapping phases: four core processes (P1: observation, P3: conceptualization, P5: operationalization, and P7: administration) and four corresponding evaluation/assessment phases (P2, P4, P6, and P8). The second dimension relates to a descriptive model of measurement error that has been articulated by Groves (1987, 1989) and modified superficially by the present author to accomplish specific goals (Esposito 2003). This model, as modified, comprises five potential sources of error: (1) questionnaire: content specialists; (2) questionnaire: design specialists; (3) interviewer; (4) respondent; and (5) mode. The framework is intended more for consideration in the design/redesign and evaluation of interviewer-administered panel surveys that have recognized and ongoing societal importance.

2.1. Questionnaire design and evaluation: An elementary process model

As noted, the process model comprises eight partially overlapping phases both for initial design and redesign efforts (see Table 1; for additional details, see Section 2.3):

2.1.1. Phase one (P1): Observation

Observation constitutes the foundation upon which science and most personal knowledge is built. During this initial phase, content specialists and other subject-matter experts focus on observable *activity* (behavior and events) within various *contexts* (family; community; workplace). While the ideal, at least initially, may be *bottom-up processing* (i.e., relatively unfiltered perception) of domain-specific behaviors and events across a broad range of contexts, it is presumed that observation by content specialists involves substantial *top-down processing* (i.e., experience- or theory-laden perception) across a more restricted range of contexts. What *survey participants* (respondents and interviewers) have observed and know is also important, because if there is a significant mismatch between what they and content specialists have observed, the design and evaluation of questionnaires is apt to be problematic.

Table 1. A framework relating questionnaire design-and-evaluation (D-and-E) processes to sources of measurement error

Questionnaire			Interdependent Sources of Measurement Error (at P7 or RP7)				
			Questionnaire D-and-E Team		Information/Data Collection Context		
			Content Specialists (1)	*Design Specialists (2)*	*Interviewer (3)*	*Respondent (4)*	*Mode (5)*
		INITIAL DESIGN					
Design and	P1	*Observation*	C_{11}: 1984	•	•	C_{14}: 1984	
Evaluation	P2	Evaluation	•	•	•	•	
Phases	P3	*Conceptualization*	C_{31}: 1984	•	•	C_{34}: 1984	
	P4	Evaluation	•	•	•	•	
	P5	*Operationalization*	C_{51}: 1984	C_{52}: 1984	•	•	C_{55}: 1984
	P6	Evaluation	•	•	•	•	•
	P7	***Administration***	C_{71}: 1984–2002	C_{72}: 1984–2002	C_{73}: 1984–2002	C_{74}: 1984–2002	C_{75}: 1984–2002
	P8	Evaluation	C_{81}: 1996–2000	C_{82}: 1996–2000	C_{83}: 1996–2000	C_{84}: 1996–2000	C_{85}: 1996–2000
		REDESIGN					
Redesign and	RP1	*Observation*	C_{R11}: 1996–2000	C_{R12}: 1996–2000	C_{R13}: 1996–2000	C_{R14}: 1996–2000	
Evaluation	RP2	Evaluation	•	•	•	•	
Phases	RP3	*Conceptualization*	C_{R31}: 1998–2000	C_{R32}: 1998–2000	•	•	
	RP4	Evaluation	C_{R41}: 1997–1998	•	•	•	•
	RP5	*Operationalization*	C_{R51}: 1999–2000	C_{R52}: 1999–2000	•	•	C_{R55}: 2000
	RP6	Evaluation	C_{R61}: 2000	C_{R62}: 2000	C_{R63}: 2000	C_{R64}: 1999–2000	C_{R65}: 2000
	RP7	***Administration***					
	RP8	Evaluation					

Note: "No activity" cells, designated with bullet symbols (•), indicate that no documented activity was conducted or recorded. Dated cells refer to documented activity that was conducted prior to, or with respect to, a specific administration of the DWS.

2.1.2. Phase two (P2): Assessment of observation phase

The concern here has to do with strong influence of prior experience and knowledge in framing and potentially distorting observations in the present. Two lines of evaluation work might be useful here: The first would assess the observation-based knowledge of questionnaire content specialists, and the second would assess the range of observation-based knowledge possessed by individuals who share the characteristics of likely survey participants. Because of their expertise as observers of social behavior in various cultural contexts, ethnographers would appear to be in the best position to perform this sort of observational assessment (e.g., Glaser and Strauss 1967/1999; cf. Webb et al. 1966). With their work as the standard, ethnographic researchers could be asked to compare their observations against those of content specialists and against those of prospective survey participants – the two lines of research alluded to above – noting significant disparities and trying to determine the origins of those disparities.

2.1.3. Phase three (P3): Conceptualization

During this phase, the *domain of interest* (i.e., the relevant "world" under investigation) is selectively abstracted and organized into a network of concepts and categories. While the capacity to form concept-based categories appears to be a universal human attribute, there is still considerable debate as to how these concepts and categories are represented in memory (Barsalou 1992; Smith 1989). Presumably, content specialists will differ with respect to which concepts and categories they identify as central and with respect to the delineation of causal interrelationships among them. An important consideration here is who assumes primary responsibility for conceptualization (e.g., an individual content specialist versus an interdisciplinary team) and how these tasks are accomplished (e.g., limited versus comprehensive domain-specific observations; discipline-specific versus interdisciplinary theoretical frameworks). Another important consideration has to do with the degree of correspondence between a sponsor's or content specialist's conceptual understanding of the target domain and that of prospective survey participants, and how well both sets of understandings reflect what "actually exists and takes place" in the target domain. (For readers with an interest in these issues, Hox (1997) provides a scholarly discussion on the topics of conceptualization and operationalization. While the approaches he describes for formulating survey questions are appealing in a theoretical sense, my limited experience in this area suggests that the approach used to generate questions for large-scale governmental surveys tends to be more empirical and pragmatic.)

2.1.4. Phase four (P4): Assessment of conceptualization phase

The inclusion of this phase is based on the belief that, as compared with the perception of physical objects, the conceptualization of social reality is to a much greater extent subject to personal, professional and cultural influences. In an effort to correct for potential personal (experiential), professional (theoretical) and cultural predilections, qualitative research – ethnographic, psychological or domain-specific – should be undertaken by skilled and independent professional observers to evaluate and reconcile the conceptual terms, models and assumptions of subject-matter experts (e.g., Gerber 1999; Miller 2002).

2.1.5. Phase five (P5): Operationalization

After a decision has been made to gather data by means of a questionnaire, content specialists and design specialists assume responsibility for translating survey concepts into questionnaire items and ancillary *metadata* (e.g., conceptual definitions and specifications; interviewing manuals; classification algorithms; see Dippo and Sundgren 2000). The development, dissemination and comprehension of survey-relevant metadata are crucial for understanding the origins of measurement error and the interrelationships between its various sources.

2.1.6. Phase six (P6): Assessment of the operationalization phase (Pretesting phase)

During this phase, design specialists – ideally in close collaboration with content specialists and field operations staff – assume primary responsibility for developing a plan to formally test the draft questionnaire. This testing usually starts with an assessment of how research participants "process" questionnaire content cognitively (i.e., comprehension, retrieval, judgment, response/reporting; see Tourangeau, Rips, and Rasinski 2000) and may involve subsequent field testing (DeMaio et al. 1993; Akkerboom and Dehue 1997). The influence of other psychological states (e.g., motivational and emotional) on the nature of the response process may or may not be considered at this point (e.g., see Cannell, Miller, and Oksenberg 1981).

2.1.7. Phase seven (P7): Survey administration

After pretesting work is completed, which could involve several P1-P6 iterations, and after modifications have been made to the questionnaire and to its pertinent metadata, the survey instrument is finalized and moved to a production environment. The administrative phase represents the locus of measurement error.

2.1.8. Phase eight (P8): Assessment of survey administration phase (Quality assessment phase)

Depending on available resources, the importance of a survey's data products (e.g., poverty and crime statistics; unemployment data) and the rate of change within the domain of interest, the sponsor may choose to conduct post-implementation quality assessment research. Virtually any of the techniques used to pretest a draft questionnaire can be used periodically to evaluate whether questionnaire items are adequately capturing and measuring the concepts specified by the survey sponsors (e.g., see Esposito and Rothgeb 1997; Forsyth and Lessler 1991).

While *social, technological and cultural change* complicates all forms of recurring social measurement, the *rate of change* that occurs in various content domains can vary widely. Given a modest rate of change associated with the content of a given social survey, one or more redesign efforts can be expected. Design and redesign processes overlap to the extent that content and design specialists make use of quality assessment findings (P8) in their redesign work (RP1 through RP6).

2.2. Sources of survey measurement error

The framework's second component involves five interdependent sources of measurement error. Groves defines *measurement error* as "the discrepancy between respondents'

attributes and their survey responses" (1987, p. S162) and distinguishes among four sources of measurement error: the interviewer, the respondent, the questionnaire, and the mode of data collection (1987, pp. S163-S166; 1989, Chapters 8 through 11). In describing measurement error arising from the questionnaire, we will find it useful to distinguish between the contributions of two specialized groups: *content specialists* (i.e., subject-matter experts with program and/or survey development responsibilities) and *design specialists* (i.e., survey practitioners/professionals who design and evaluate questionnaires, prepare training materials, develop algorithms, etc). The rationale for this distinction is rooted in the different roles each group assumes in the questionnaire design-and-evaluation process and in the specialized expertise each possesses with regard to resolving certain types of issues and problems (theoretical/conceptual versus technical design). From a functional perspective, content and design specialists, as an integrated working group, constitute the *questionnaire design-and-evaluation team*; the setting that incorporates the interviewer, the respondent and the collection mode constitutes the *information/data-collection context*. Brief descriptions of the five sources of measurement error are provided below.

2.2.1. Questionnaire: Content specialists

Especially during the observation and conceptualization phases associated with questionnaire design, content specialists assume a central role in describing the domain of interest, isolating and defining key concepts and categories, and delineating possible relationships among theoretical variables (Federal Committee on Statistical Methodology 1988; Hox 1997; Turner and Martin, 1984, Chapter 7). Their assumptions and theories be they explicit or implicit, about how domains are structured, about how theoretical relationships change over time, and about why actors behave as they do in various situations, have a profound effect on questionnaire content and data quality. The more "accurate" their observations, concepts, assumptions and theories, the more successful the survey measurement process is likely to be.

2.2.2. Questionnaire: Design specialists

During initial design, questionnaire-design specialists, usually following guidelines prescribed by researchers and practitioners (Belson 1981; Converse and Presser 1986; Foddy 1993; Fowler 1995; Sudman and Bradburn 1982), transform conceptual specifications provided by content specialists into coherent sets of questionnaire items and ancillary metadata. Even when conceptual specifications appear reasonably clear and precise, this translation/design process can be challenging.

2.2.3. Interviewers

With respect to minimizing measurement error, there would appear to be disparate views among researchers and practitioners as to the proper role of interviewers in administering surveys (Beatty 1995; Maynard and Schaeffer 2002). For some, their prescribed role is to administer survey questions in a standardized manner (Fowler and Mangione 1990). For others, their prescribed role is to facilitate the communication of intended "meaning" when administering survey questions (Suchman and Jordan 1990), which may require

a more flexible approach to asking questions and providing feedback (Conrad and Schober 2000). Since neither prescribed role can be expected to remove interviewers as a potential source of measurement error, survey sponsors need to consider the relative costs and benefits associated with efforts to do so. Whatever one's position on this issue, content and design specialists would be wise to resist the temptation to reflexively assign blame to interviewers for questionnaire-administration problems that, on closer inspection, might be found to have their locus in early design-and-evaluation work (e.g., P1, P3, and P5).

2.2.4. Respondents

In an effort to improve data quality, behavioral scientists: (1) have developed socio-cognitive models of the response process (Cannell, Miller, and Oksenberg 1981; Tourangeau 1984; for a review, see Jobe and Herrmann 1996); (2) have described the types of cognitive errors that can occur at each stage (Tourangeau, Rips, and Rasinski 2000); and (3) have devised strategies for identifying questionnaire problems and reducing measurement error (Schwarz and Sudman 1996; Gerber 1999). Considerable gains appear to have been made in exploiting cognitive strategies to reduce error (Sirken et al. 1999; Jobe and Mingay 1989; cf. O'Muircheartaigh 1999). Sometimes, however, problems with the response process may be traced to a significant motivational component (e.g., content irrelevance; competing time demands). When unmotivated to participate fully in a survey, respondents may engage in satisficing behavior (Krosnick 1991), thus increasing their contribution to the magnitude of measurement error.

2.2.5. Mode

The selection of a data-collection mode (or modes, as the case may be) clearly has an effect on estimates of measurement error (Tourangeau, Rips, and Rasinski 2000, pp. 289–312; cf. Groves 1989, pp. 501–552). Oftentimes, the choice of mode is dictated by cost considerations, and modest increases in measurement error tend to be accepted as part of the compromise to reduce survey costs.

2.3. Additional details regarding the framework (Table 1)

Several additional aspects of the framework are worthy of note. First, it is presumed that design-and-evaluation work can and often does overlap across phases and that movement between certain phases (P1 through P6) is bidirectional and potentially iterative. Second, the phrase "interdependent sources of measurement error" has been adopted to reflect the view that measurement error – and accuracy, too – is presumed to be the outcome of collaborative or interactive processes involving the various sources of error identified in Table 1 (Suchman and Jordan 1990, pp. 240–241). Within a given data-collection context, measurement error is presumed to be a byproduct of role- and task-specific activities – Sudman and Bradburn's (1974) terminology (cf., Platek 1985) – that manifest themselves during the survey administrative phase (P7 or RP7). Various role- and task-specific activities that are performed inadequately at prior design-and-evaluation phases (P1 through P6) can be viewed as *precursors* to measurement error. Third, the actual performance of role- and task-specific activities, represented as generically labeled cell entries (e.g., C_{12}), is presumed to vary across survey design-and-evaluation efforts.

Whether a particular cell has an entry or not would depend on whether specific cell-related activities were conducted and whether documentation exists for those activities. For example, if content specialists are not involved in pretesting work conducted during the initial questionnaire design, then cell C_{61} would be left blank. Empty cells are problematic in that they represent activity or knowledge gaps that are apt to affect the locus and magnitude of measurement error. And lastly, as noted, social, technological and cultural change also plays a crucial role in the measurement process. Unless continuously monitored and accounted for by content and design specialists, rapid change within a given target domain can have a substantial effect on the magnitude of measurement error.

3. Target Questionnaire: The Displaced Worker Survey/Supplement

3.1. Brief history

In the early 1980s, the American economy was staggered by two recessions that were especially hard on manufacturing industries, particularly steel and automobile production. In an effort to assess the effects of these developments on the labor force, a small group of labor economists (content specialists) at the U.S. Bureau of Labor Statistics, in collaboration with design specialists at the U.S. Census Bureau, set about to design a questionnaire to be used in a survey that would estimate the number of workers who were displaced from jobs (Table 1, Cells C_{51} and C_{52}). This survey, known to data users as the Displaced Worker Survey (DWS), was first administered as a supplement to the Current Population Survey (CPS) in 1984. Although the DWS was intended to be a one-time survey (administered in 1984 only), the data it generated had utility for both internal and external users and, as a result, it has been administered biennially ever since. The primary objective of the supplement is to estimate the number of workers who have lost or left a job for specified displacement reasons and to collect data on the types of jobs that these workers have lost or left (e.g., industry, occupation, earnings). "While there never has been a precise definition for (*displaced workers*), the term is generally applied to persons who have lost jobs in which they had a considerable investment in terms of tenure and skill development and for whom the prospects of reemployment in similar jobs are rather dim" (Flaim and Sehgal 1985, p. 4).

3.2. In-house review of the DWS questionnaire

In June 1995, the present author was asked to review the DWS to identify potential sources of measurement error. The review, which was not based on a formal coding scheme, but which was sufficient for the purposes intended, identified a number of potential problems with the DWS questionnaire: (1) problematic question wording, especially with respect to two key items (SD1 and SD2, see below); (2) ambiguous conceptual terminology; and (3) unclear or incomplete question specifications. Concern about these problems prompted the supplement sponsors to authorize that quality assessment research be conducted in February 1996.

3.3. Key supplement questions: SD1 and SD2

Most of the evaluation data to be reviewed in this article focuses on two key supplement items: SD1 and SD2 (see Table 2 for item wording and skip instructions). The reason for focusing on these items is that they carry most of the burden for classifying workers who have separated from jobs during the reference period as displaced or not displaced. From an analytical perspective, an understanding of the metadata associated with these items is indispensable for detecting evidence of measurement error (see Esposito 2002, Appendix, for relevant DWS metadata).

4. Methodology

The research conducted on the displaced-worker supplement during the period 1995–2000 is based on a multiple-method approach to questionnaire evaluation that was used in the early 1990s by researchers at the BLS and the U.S. Census Bureau to redesign the CPS

Table 2. Supplement items SD1 and SD2 (Adults, unweighted data, 1996–2000)

1996 [*N*=76,112]	**1998** [*N*=79,503]	**2000** [*N*=79,121]	**SD1.** During the last 3 calendar years, that is January (1993/1995/1997) through December (1995/1997/1999), did you lose a job or leave one because: Your plant or company closed or moved, your position or shift was abolished, insufficient work, or another similar reason?
8.9%	7.3%	7.4%	⟨1⟩ Yes (Go to SD2)
91.1%	92.7%	92.6%	⟨2⟩ No (End Displacement Series)
			SD2. Which of these specific reasons describes why you are no longer working at that job?
1996 [*N*=6608]	**1998** [*N*=5838]	**2000** [*N*=5854]	READ IF NECESSARY: If you lost or left more than one job in the last 3 years, refer to the job you had the longest when answering this question and the ones to follow. [Note: *Interviewers are instructed to read all six response options.*]
22.2%	24.5%	23.4%	⟨1⟩ Plant or company closed down or moved
			Plant or company still operating but lost or left job because of:
26.4%	22.0%	20.2%	⟨2⟩ Insufficient work
15.8%	16.4%	14.0%	⟨3⟩ Position or shift abolished
4.1%	4.8%	4.3%	⟨4⟩ Seasonal job completed
1.5%	1.4%	1.5%	⟨5⟩ Self-operated business failed
29.9%	31.0%	36.6%	⟨6⟩ Some other reason
			[Skip Instructions: Precodes 1–3 proceed with the next question in the series; precodes 4–6 are skipped around the displacement series.]

(Rothgeb et al. 1991). Evaluative research methods are used to gather qualitative and quantitative data about various aspects of the survey measurement process (e.g., the interpretation of key concepts, the comprehension of question meaning, the efficiency of interviewer-respondent interactions). Data gleaned from multiple methods can be combined and contrasted to provide researchers with a more comprehensive picture of how well target questions are meeting their stated objectives (Cannell et al. 1989; Oksenberg, Cannell, and Kalton 1991; Sykes and Morton-Williams 1987).

4.1. Principal evaluation methods

As noted, three principal evaluation methods were used during each phase of this multiphase research effort: (1) interviewer debriefing; (2) interaction/behavior coding; and (3) respondent debriefing. The rationale for the repeated use of these three methods is as follows. First, collectively, the three general methods capture or reveal the perspectives of the various parties involved in the survey measurement process – interviewers, respondents, content and design specialists (see Section 6.2). Second, certain members of the research team had used these methods in prior research efforts (Esposito and Rothgeb 1997) and they had been found to be efficient, effective and relatively inexpensive. And third, to maintain a level of methodological comparability across phases, we wanted the replications to be as uniform as possible.

4.1.1. Interviewer debriefing

While there are a variety of ways to gather evaluative information from interviewers (Converse and Schuman 1974; DeMaio 1983; DeMaio et al. 1993), we debriefed interviewers using a focus group format. During the phase two evaluation, we also incorporated a target-question rating form. In an effort to minimize cost, debriefing sessions were conducted with CPS interviewers who worked at one or more of the U.S. Census Bureau's three telephone centers. Several days prior to administering the DWS, interviewers selected to participate in the focus groups were given *log forms* on which to record any problems they might have experienced with target questions. The purpose of these debriefing sessions was to obtain feedback from interviewers regarding the performance of target questions (i.e., SD1 and SD2, specifically, and, in phase three, respondent debriefing items). An extensive protocol of probe questions was used to guide the group discussion and stimulate interviewer feedback. Focus group sessions were audiotaped and written summaries were prepared from these tapes.

4.1.2. Interaction/behavior coding

Behavior coding – a specific type of interaction coding – involves a set of procedures which have been found useful in identifying problematic questionnaire items (Cannell and Oksenberg 1988; Esposito, Rothgeb, and Campanelli 1994; Fowler 1992; Fowler and Cannell 1996; Morton-Williams 1979; Morton-Williams and Sykes 1984; Oksenberg, Cannell, and Kalton 1991; Shepard and Vincent 1991). The coding form used in this research effort included six *interviewer codes* (exact reading; minor change; major change; probe; verify; and feedback) and eight *respondent codes* (adequate answer;

qualified answer; inadequate answer; request for clarification; interruption; don't know; refusal; and other).

Behavior coding was conducted at one or more of the U.S. Census Bureau's three telephone centers using a paper-and-pencil coding form and it was done live, that is, while the interview was in progress. The present author monitored CPS interviews from a supervisor's station (out of view for interviewers), selected cases to code, and coded interactions between interviewers and respondents during supplement administration. For a particular item, only data from the first exchange between the interviewer and respondent was analyzed; at either end of an exchange (interviewer side; respondent side), a maximum of two behavior codes was assigned. Extended interactions were coded, when possible, for key supplement items.

4.1.3. Respondent debriefing

While there are various techniques available for gathering evaluative information/data from survey respondents (Belson 1981; DeMaio et al. 1993; Forsyth and Lessler 1991), we used *response-dependent follow-up probing* (also see Campanelli, Martin, and Creighton 1989; Campanelli, Martin, and Rothgeb 1991; Hess and Singer 1995; Oksenberg, Cannell, and Kalton 1991; cf. Schuman 1966). A small interdisciplinary team of design and content specialists drafted the respondent debriefing questionnaire. The total number of debriefing questions varied from one phase to the next. The debriefing items were designed: (1) to gather job-related information that was relevant to job separation concepts, and (2) to determine whether item-specific problems existed that might jeopardize an accurate count of displaced workers. Each debriefing question was designed with a specific objective in mind. Answers to debriefing questions were very useful in helping the research team to detect potential sources of measurement error. To minimize cost and respondent burden, the research team restricted respondent debriefing to approximately 25 percent of the CPS sample, about 13,000 households. The sequencing of questions went as follows: Respondents were first asked the basic CPS questions for all eligible household members, then supplement questions for all eligible household members, and then the debriefing questions. Certain demographic and labor force criteria determined which displacement questions the respondent was eligible to be asked. These criteria, and responses to specific supplement items, determined which debriefing questions the respondent was asked.

Having provided a description of the general methodology for this multiphase effort, let us now turn to a discussion of the three phases of evaluation research (for an overview, see Table 3).

5. Methodological Details, Findings, Discussion, and Implications

5.1. The first evaluation: 1996

In retrospect, this initial evaluation can best be described as *exploratory* quality assessment research. The research plan, a collaborative effort involving BLS and Census Bureau personnel, was implemented by field staff and two behavioral scientists (Esposito and Fisher 1998).

Table 3. Overview of methods and findings for three evaluation phases (1996–2000)

	Comments (**C**), Methodological details (**D**) and Illustrative findings (**F**)
Phase 1 (1996)	**C:** This phase can best be described as exploratory quality assessment research. This initial evaluation focused on two supplement items, SD1 and SD2.
Interviewer debriefing	**D:** One focus group involving 10 telephone center interviewers. **F:** Evidence of conceptual problems (e.g., what constitutes a job), cognitive problems (e.g., meaning of the phrase "or another similar reason"; difficulty with the distinction between losing and leaving a job) and design/operational problems (e.g., failure to read all parts of questions).
Interaction coding	**D:** 52 person interviews coded (behavior coding). **F:** Evidence of problems with interviewers reading SD1 and SD2 as worded (12% and 57% of cases with major changes, respectively); respondents also had difficulty providing adequate answers to SD2 (33% of cases had inadequate answers).
Respondent debriefing	**D:** Debriefing questionnaire consisting of 8 response-dependent probe questions. **F:** Evidence of possible displaced-worker undercount in the order of 25 percent (false negatives). About one-third of the suspected undercount was traceable to SD1, precode 6, and the remainder to inaccurate "no" answers to SD1 (unexplained).
Phase 2 (1998)	**C:** Relative to the quality assessment work conducted in 1996, this second phase was far more comprehensive. Again, the evaluation focused on SD1 and SD2.
Interviewer debriefing	**D:** Three focus groups involving 34 telephone center interviewers. Interviewers were also asked to rate SD1 and SD2 in terms of how difficult they thought these items were for respondents to answer. **F:** Evidence of conceptual problems (e.g., what to do about temporary jobs and other alternative work arrangements), cognitive problems (e.g., uncertainty regarding the meaning of terms such as "insufficient work" and "layoff") and design/operational problems (e.g., awkward transition phrase in SD2; parents reporting for older children; burden on the elderly and the disabled; interruptions). Rating scale data (means and standard deviations) for SD1 and SD2 provided evidence of considerable variability within and between groups of telephone center interviewers.
Interaction coding	**D:** 145 person interviews coded (behavior coding). **F:** Evidence of problems reading SD1 and SD2 as worded (13% and 72% of cases with major changes, respectively); respondents also had difficulty providing adequate answers to both items (10% and 28% of cases had inadequate answers, respectively).

Table 3. Continued

Respondent debriefing	**D:** Debriefing questionnaire consisting of 22 response-dependent probe questions. **F:** Evidence of possible displaced-worker undercount in the order of approximately 20 percent (false negatives). Again, about one-third of the suspected undercount was traceable to SD1, precode 6, and the remainder attributable to inaccurate "no" answers to SD1 (unexplained). However, other debriefing data raises questions as to the actual status of some "displaced workers" (e.g., 23% of cases categorized as displacements due to "insufficient work" were later reported to have been temporary jobs); some labor force economists would exclude persons whose jobs were temporary from the count of displaced workers (potential false positives).
Phase 3 (2000)	**C:** This third evaluation was moderate in size and involved both quality assessment work (again, SD1 and SD2) and pretesting work (i.e., evaluated a subset of respondent debriefing items under consideration for a new, broader supplement on job separations).
Interviewer debriefing	**D:** Two focus groups involving 22 telephone center interviewers. **F:** Both supplement items and preselected debriefing items were evaluated during this phase. With respect to SD1 and SD2, some additional evidence of conceptual problems was noted (e.g., what to do about mergers and job transfers). Several respondent debriefing items, currently under consideration for a new supplement on job separations, also manifested a variety of conceptual problems (e.g., what to do about "job switching" within a company; freelance work), cognitive problems (e.g., uncertainty regarding the subtle differences between losing and leaving a job) and design/operational problems (e.g., accurately categorizing answers given a list of 20 response precodes).
Interaction coding	**D:** 131 person interviews coded (behavior coding). **F:** Again found evidence of problems reading SD1 and SD2 as worded (18% and 43% of cases with major changes, respectively); respondents also had difficulty providing adequate answers to SD2 (28% of cases had inadequate answers). Four debriefing items (SDB2A/B and SDB5A/B) that are similar to supplement item SD2 in purpose, but not format, outperformed SD2 but still proved difficult to read as worded (21% major changes, combined data); respondents struggled with these items as well (26% inadequate answers, combined data).
Respondent debriefing	**D:** Debriefing questionnaire consisting of 11 response-dependent probe questions. **F:** Evidence of a possible displaced-worker undercount of 29 percent (false negatives); however, prior work (phase two) suggests that this figure may be overstated due to the temporary nature of the jobs that were lost. In contrast to prior evaluations, which were based on a full three-year reference period (e.g., 1997–1999), this particular estimate is based on data for the most recent year (1999). Once again, about one-third of the suspected undercount was traceable to SD1, precode 6, and the remainder to inaccurate "no" answers to SD1 (unexplained).

5.1.1. Methodological details and findings

Relative to subsequent phases, the scope of this initial evaluation was limited. With respect to *interviewer debriefing*, one focus group was conducted with ten CPS interviewers serving as research participants. Interviewers mentioned a number of conceptual and operational problems associated with SD1 and SD2. A summary of findings can be found in Table 3 and examples of questions used to debrief interviewers on these two items appear in Table 4.

With regard to *behavior coding*, 23 CPS household interviews were monitored and interviewer-respondent exchanges for 52 interviews were coded. Coded data suggest that interviewers experienced some difficulty reading SD1 as worded but that respondents provided adequate answers on a fairly regular basis. Relative to SD1, item SD2 was asked much less frequently in that only a small percentage of persons lost or left jobs during the three-year reference period. As a result, these data should be interpreted with caution. Interviewers struggled when trying to read SD2 as worded and respondents experienced some difficulty in providing adequate answers (see Table 5).

With respect to *respondent debriefing*, a debriefing questionnaire of follow-up probes was developed that comprised eight items (see Table 6 for examples); sample sizes for these items ranged from $n = 66$ to $n = 17{,}605$. These debriefing items were useful in identifying and quantifying potential measurement error. For example, debriefing item SDB5 was asked of a sample of individuals who had lost/left a job during the three-year reference period but for whom their reason-for-separation was coded as "some other reason" (SD2, precode 6). The DWS classification algorithm excludes all such individuals (30 percent of all responses to SD2 in 1996) from the count of displaced workers. When SDB5 was asked, however, about 19 percent of these cases involved target persons who

Table 4. Examples of interviewer debriefing questions (Phase one, 1996)

SD1	Did you have difficulty reading this question in its entirety before respondents provided an answer? Did the respondents have difficulty with the concept of "lose a job or leave one"? Were respondents able to distinguish the four response options presented to them? If not, what confusions or misconceptions did they report? Was the phrase "or another similar reason" causing any problems for respondents? How clear were instructions on classifying a response so it could be matched to one of these four options?
SD2	Did you have difficulty reading this question in its entirety (i.e., all 6 response options)? Did the list of reasons (1–5) seem to cover most respondents or did a large percentage of respondents get coded into "some other reason"? How frequently did you read the READ AS NECESSARY statement? Did respondents understand the meaning of each of the reasons provided for their nonemployment? If not, which reasons did respondents fail to understand? And why? Were there any additional reasons offered by respondents for their job loss not available in the current list? If yes, what were they?

Table 5. Behavior coding data for selected items (1996–2000)

Phase	Item(s)	Interviewer codes		Respondent codes			
		E	MC	AA	IA	RC	INT
One (1996)	SD1	65% (33/51)	16% (8/51)	88% (42/48)	2% (1/48)	8% (4/48)	19% (9/48)
	SD2	29% (2/7)	57% (4/7)	67% (4/6)	33% (2/6)	0% –	17% (1/6)
Two (1998)	SD1	71% (96/135)	13% (18/135)	88% (119/135)	10% (13/135)	1% (1/135)	25% (34/135)
	SD2	0% –	72% (13/18)	56% (10/18)	28% (5/18)	0% –	39% (7/18)
Three (2000)	SD1	69% (82/119)	18% (22/119)	93% (110/118)	5% (6/118)	0% –	13% (15/118)
	SD2	29% (4/14)	43% (6/14)	60% (6/10)	40% (4/10)	0% –	0% –
	SDB3	93% (110/118)	3% (4/118)	98% (115/117)	0% –	2% (2/117)	5% (6/117)
	[SDB2A/B + SDB5A/B]	74% (14/19)	21% (4/19)	74% (14/19)	26% (5/19)	0% –	16% (3/19)

Notes. Data are presented for key supplement and debriefing questions (**SD** and **SDB** prefixes, respectively) and only for the most informative interviewer and respondent codes. Codes may sum to a value larger than 100% because a maximum of two codes is permitted on both sides of an exchange. Ratios (*c*/*n*) refer to the number of times a code was assigned (*c*) divided by the number of times the question was asked (*n*). Also, given the limited number of times SBD2A/B and SDB5A/B were administered, data for these items were combined.

Abbreviations. Interviewer codes: E (exact reading) and MC (major change in wording). Respondent codes: AA (adequate answer), IA (inadequate answer), RC (request for clarification), and INT (interruption).

Table 6. Examples of respondent debriefing questions (Phase one, 1996)

SDB1	Earlier you told me that you had lost or left a job during the past three calendar years. Did you lose or leave more than one job in the time period spanning January 1993 through December 1995?
	Rationale: The DWS had no explicit mechanism for identifying persons who lost or left more than one job during the reference period. This is a problem because the DWS only collects data for one job and, in such cases, respondents need guidance on which job to report (see SDB2).
SDB2	Earlier in this interview, when answering questions about the job you had lost or left from January 1993 through December 1995, were you answering the questions based on the job that you had held for the longest time?
	Rationale: Since persons who lost or left more than one job were not explicitly told on which job to report (i.e., the longest held job *from which a displacement occurred*), reporting errors were possible. SDB2 was an attempt to quantify that error.
SDB3	Did you lose that job or did you leave that job?
	Rationale: In this context, SDB3 is best classified as an informational probe. The supplement sponsor wished to know what percentage of displaced workers had lost a job relative to those who had left a job.
SDB4	During the time period spanning January 1993 through December 1995, did you leave a job or retire from a job?
	Rationale: SDB4 was asked of all persons for whom a "no" answer was provided to supplement item SD1. The goal was to identify persons who might have been missed as displaced workers (see SDB5).
SDB5	What was the exact reason (you/he/she) (are/is) no longer working at that job? (Note: Eighteen substantive response precodes were provided, eight of which described displacement scenarios (e.g., *company or plant* had insufficient work; was downsizing or restructuring; was filing for bankruptcy).)
	Rationale: SDB5 was asked of all persons for whom a "yes" response was given to debriefing item SDB4. If the response to SDB5 matched one of the eight displacement precodes, that case was classified as a potential false negative.

had indeed lost/left a job for a displacement reason – 84 cases, all possibly false negatives, from this one path alone. A second path, persons for whom a "no" answer was provided initially in the case of SD1 but for whom responses to subsequent debriefing questions (SDB4 and SDB5) suggested that they may have been displaced, yielded an even higher number of potential false negatives, 174 cases. When the number of false negatives for each path is adjusted for the 25% debriefing-question sampling rate (i.e., multiplied by four), then combined ($336 + 696 = 1{,}032$, numerator), and then divided by the appropriate denominator (i.e., base equals 4,211, the sum of precodes 1, 2, and 3 for SD2), these debriefing data suggest a displaced-worker undercount of approximately 25 percent. As can be seen, about one-third of this error is traceable to path one (SD2, precode 6), and the remainder to path two (SD1 = no).

5.1.2. Discussion and implications for subsequent evaluations

This initial evaluation provided both quantitative and qualitative evidence of problems with supplement items SD1 and SD2. Behavior-coding data suggested that these two items are difficult for interviewers to read and for some respondents to answer. Respondent debriefing data suggested that design problems with SD1 and SD2 might have led to a substantial undercount of displaced workers, perhaps as much as 25 percent. And qualitative data generated during the focus group corroborated some of the findings noted above and raised other concerns about conceptual issues. All three sources of evaluation data seemed to converge on the conclusion that SD1 and SD2 were flawed and that a substantial amount of measurement error was being generated as a result. The remedy seemed obvious: Commence work on a redesign of the DWS. Due in part to concerns about what we still did not know, a redesign effort was not undertaken at that time. In retrospect, this proved to be a wise decision, because while we had learned much, there was still much left to learn.

The conceptualization problems raised in this first evaluation prompted a review of supplement metadata and stimulated discussion among internal content specialists as to what they understood a displaced worker to be. This review and discussion produced some interesting revelations and insights. First, there was relatively little documentation available on the initial conceptualization process or on the observations that inspired the concept. Second, concept specifications were not always explicitly operationalized or were implemented in counterintuitive ways. For example, interviewer instructions state that persons laid off from a job are to be counted among the displaced if recalled to a job (e.g., assembler) different from the one from which they were laid off (e.g., welder); however, there are no questions in the supplement that address this specific issue. Regarding counterintuitive implementations, the phrase "or another similar reason" in SD1 would seem to mean reasons *similar to one of the (displacement) reasons explicitly stated in the question*. The DWS instructional memorandum defines the phrase as follows: "These include all types of factors which are based on the operating decisions of the firm, plant or business in which the worker was employed and which result in the worker losing or leaving a job"(U.S. Bureau of the Census 2000, p. 5). Though somewhat vague, the information provided to interviewers appears to confirm our lay impression; however, all such cases are skipped out of the displacement series (see SD2, precode 6). (Note: Though counterintuitive, the decision to skip "some other reason" entries out of the displacement series actually *reduced* measurement error. Including those cases would have generated about four times as many false positives (81%) as false negatives (19%). Most of the precode-6 entries do not constitute displacements.) Third, various aspects of the displacement concept, as understood by content specialists (most of whom had not been involved in the original design work), had apparently changed over the years in ways that the supplement was not designed to measure – we might call this "conceptual drift." For example, whereas the supplement makes no substantive distinction between persons who *lose jobs* and those who *leave jobs* for displacement reasons, current thinking is that persons in the latter group probably should be required to satisfy additional conditions to be classified as displaced (e.g., written notification of impending job loss). And lastly, the *domain of interest* (i.e., actual manifestations of displacement in the observable world of work) had also changed, and this, too, had created measurement problems. For example,

short-term work arranged through temporary staffing agencies had become much more common over the intervening 20 years precipitating debate among content specialists as to how such work arrangements should be handled.

The issues and problems noted above regarding observation, conceptualization, and operationalization relate directly to the first three core phases of the design-and-evaluation process described earlier (see Table 1, Phases P1, P3, and P5) and provide a sense of how such problems contribute to measurement error during the survey administration phase (P7). The intent of pretesting (P6) is to identify and remedy such problems, but if evaluation work is poorly designed or superficial, such problems may go undetected or they may be misrepresented. To my knowledge, there was no formal pretesting work conducted on the DWS. And even though the quality assessment work (P8) conducted in 1996 provided direct and indirect evidence of measurement error (at P7), one always needs to exercise care in interpreting evaluation findings. For example, after reading a draft of the evaluation report, one reviewer, a subject-matter specialist, noted that very little effort had been expended in identifying false positives – a valid observation. An attempt to correct this imbalance was made in subsequent research. Also, small-scale evaluations limit the numbers of individuals who provide information and data and, as a result, lead to questions of representativeness and thoroughness. Conducting a single focus group with telephone-center interviewers (but not other field-based interviewers) represents a potential source of *evaluation error* (i.e., misleading or inaccurate information/data collected during a specific evaluation effort). Relying on a single researcher to conduct behavior coding represents another source of evaluation error. And decision-making regarding the number, content and design of respondent debriefing questions can represent yet another source of evaluation error. Had any one of these evaluation methods been used alone, there would have been cause for concern regarding the utility of research findings. However, a multiple-method evaluation strategy, with its inherent checks and balances, provides researchers with some degree of protection against serious single-method evaluation error and helps to allay fears that a significant source of evaluation error will undermine research findings.

5.2. The second evaluation: 1998

Fortified with metadata from the 1996 evaluation, issues that had not occurred to the research team prior to the first evaluation were now apparent and open for discussion. A formal working group of content and design specialists met regularly to review research findings, discuss conceptual issues, and formulate plans for a larger, more comprehensive evaluation.

5.2.1. Methodological details and findings

This second evaluation substantially expanded the scope of inquiry. With respect to the interviewer debriefing, three focus groups were conducted with 34 CPS interviewers serving as research participants – one at each of the three centralized telephone facilities. The questions used to debrief interviewers on SD1 and SD2 were virtually identical to those asked in the prior phase (Table 4). Most of the substantive problems identified during phase one were again observed here (see Table 3). Many of the problems noted

by interviewers increase the likelihood of *categorization errors* (i.e., not checking the best precode or the correct precode) and *classification errors* (i.e., categorization errors that result in persons being misclassified as displaced or not displaced – false positives and false negatives, respectively).

In addition to gathering qualitative information about SD1 and SD2 during the debriefing sessions, we asked interviewers to rate these items in terms of how difficult they thought it was for respondents to provide adequate answers (see Table 7). The goal was to obtain a crude sense of the frequency of problems experienced with these items. As can be seen, SD1 was identified as problematic in all three sessions; and means (1.67, 2.20, 2.67) and individual ratings differed considerably between and within groups. Somewhat surprisingly – given focus group discussions and the magnitude of the ratings for this item – SD2 was only identified as problematic by two of the groups; and, again, means (2.00, 3.00) and individual ratings varied considerably between and within groups.

Like the debriefing of interviewers, we also expanded the collection of behavior-coding data. Sixty-three household interviews were monitored at two centralized telephone facilities and interviewer-respondent exchanges for 145 person interviews were coded. Much as in the first phase, interviewers struggled with the wording of both SD1 and SD2, especially the latter, and respondents experienced difficulties providing an adequate answer to SD2 (see Table 5).

With respect to respondent debriefing, a debriefing questionnaire of follow-up probes was developed that comprised 22 unique items (for examples and rationales, see Table 8). Sample sizes for these items ranged from $n = 4$ to $n = 18{,}477$. As was the case in phase one, debriefing items were useful in identifying and quantifying potential measurement error. For example, SDB3 was asked of a sample of persons who lost/left jobs during the three-year reference period but for whom "some other reason" was entered as the separation reason (SD2, precode 6) – about 31 percent of the responses to SD2. In about 16 percent of these cases, during the debriefing, respondents indicated that the target person had indeed lost/left a job for a displacement reason (e.g., downsizing, restructuring; position/shift abolished) – a total of 57 cases, all possibly false negatives, from this one path alone. A second path, persons for whom a "no" answer was provided initially in the case of SD1 but for whom responses to subsequent debriefing questions (SDB17 and SDB20) suggested that they might have been displaced, yielded an even higher number of potential false negatives, 129 cases. When the number of false negatives for each path is adjusted for the 25% debriefing-question sampling rate (i.e., multiplied by four), then combined ($228 + 516 = 744$, numerator), and then divided by the appropriate denominator (i.e., base equals 3,670, the sum of precodes 1, 2, and 3 for SD2), these debriefing data suggest a displaced-worker undercount of approximately 20 percent. As the data suggest, almost a third of this error is traceable to path one (SD2, precode 6), and the remainder is attributable to path two (SD1 = no).

These data corroborate findings from phase one regarding a possible undercount of displaced workers (false negatives), but other debriefing data raise questions about potential false positives. For example, data from debriefing item SDB2 indicate that approximately 23 percent of persons classified as displaced because of "insufficient work" were reported to have been working at a temporary job; the corresponding percentages for "plant or company shut down" and for "position or shift abolished" were 6.0 percent

Table 7. Interviewer ratings for supplement items SD1 and SD2 (Phase two, 1998)

Location		Interviewer ratings												Mean	*SD*
TTC	SD1:	3	1	2	2	3	1	1	1	1	1	1	3	1.67	0.89
Tucson	SD2:	–	–	–	–	–	–	–	–	–	–	–	–	–	–
HTC	SD1:	3	2	2	2	3	1	2	2	1	4			2.20	0.92
Hagerstown	SD2:	3	3	1	1	4	1	1	2	2	2			2.00	1.05
JTC	SD1:	4	2	2	2	4	4	4	1	3	2	2	2	2.67	1.07
Jeffersonville	SD2:	3	2	3	2	4	1	3	4	5	3	3	3	3.00	1.04

Note: Interviewers were asked to rate problematic supplement items using the following evaluation scale: *Based on your experiences this past week, how frequently have respondents had difficulty providing an adequate answer to (the target question) when asked?*

A (1). *Never or Very Rarely (0 to 5% of the time)*
B (2). *Occasionally (some % in between A and C)*
C (3). *About Half of the Time (approximately 45–55% of the time)*
D (4). *A Good Deal of the Time (some % in between C and E)*
E (5). *Always or Almost Always (95 to 100% of the time)*

Table 8. Examples of respondent debriefing questions (Phase two, 1998)

SDB1	Earlier you told me that me that you had lost or left a job in the past three calendar years because (*fill with displacement reason from SD2*). Did you lose that job or did you leave that job? *Rationale:* The supplement sponsor wished to know what percentage of displaced workers had lost a job relative to those who had left a job. We presumed the respondent could make this distinction without guidance from the sponsor. This probe also is used to channel job leavers to specific follow-up probes.
SDB2	Was the job you (fill as appropriate: "lost", "left", or "retired from") a temporary job, that is, a job that was supposed to last only for a limited time or until the completion of a project? *Rationale:* To identify persons whose jobs were not considered "permanent." Though the DWS does not identify such workers, persons who lose or leave temporary jobs probably should not be counted among the displaced.
SDB3	Some people leave jobs for personal reasons, such as to further their education or to care for children. Others lose or leave jobs for economic reasons, such as insufficient work or downsizing. What is the MAIN reason you are no longer working at that job? (Note: This item had 22 response precodes, seven employer-related reasons (e.g., business closed down; restructuring; insufficient work; position/shift abolished) and fifteen personal reasons (e.g., did not like job or boss; better job; not enough pay; own illness/injury; fired; school/training). *Rationale:* Generally speaking, to determine if the person lost or left a job involuntarily (i.e., one of the employer-related reasons) or voluntarily (i.e., one of the personal reasons). With respect to employer-related reasons only, this item was useful for identifying potential false negatives.
SDB3Z	Did you ever return to work for that employer, for even a short period of time? *Rationale:* For persons reported to have lost, left, or retired from a job during the reference period for a displacement reason, to determine if the person returned to work for that employer, even briefly. This item is an attempt to identify individuals who might be considered false positives (e.g., persons who returned to work for their former employers, presumably doing the same work and not subsequently displaced again).
SDB17	During the period January 1995 through December 1997, did you leave a job or lose a job for any reason? *Rationale:* SDB17 was asked of all persons for whom a "no" answer was provided to supplement item SD1. The goal was to identify persons who might have been missed as displaced workers (see SDB20).
SDB20	What is the MAIN reason you are no longer working at that job? (Note: This item had 22 response precodes, seven employer-related reasons) and fifteen personal reasons (see SDB3 for examples). *Rationale:* Generally speaking, to determine if the person lost or left a job involuntarily (i.e., one of the employer-related reasons) or voluntarily (i.e., one of the personal reasons). With respect to employer-related reasons only, this item was useful for identifying potential false negatives.

and 11.5 percent, respectively. Some subject-matter specialists would argue that temporary workers should not be counted among the displaced, regardless of the reason for separation. Data from another debriefing question, SDB3Z, yielded a similar pattern: approximately 13 percent of persons classified as displaced because of "insufficient work" were reported to have returned to work for their former employers, however briefly; the corresponding percentages for "plant or company shut down" and for "position or shift abolished" were 3.6 percent and 11.9 percent, respectively. Some of these persons may have been misclassified as displaced workers.

5.2.3. Discussion and implications for subsequent research

This second evaluation provided a wealth of qualitative and quantitative data, and was important in two respects. First, as a partial replication of the first evaluation, it was successful in corroborating prior findings and convincing program managers and content specialists that the problems identified earlier were real. Moreover, rating-form data provided a crude measure of how much difficulty respondents (and interviewers) were experiencing with items SD1 and SD2. Behavior-coding data collected during this phase proved to be consistent with data collected during phase one, and quantified the difficulties that interviewers and respondents experience in asking and responding to these items. And the respondent debriefing questionnaire again generated quantitative evidence that pointed to a significant amount of measurement error (i.e., about 20 percent false negatives). Secondly, this phase was important in that it provided a more balanced view of the measurement error associated with the DWS. Phase one was useful in detecting false negatives; phase two was successful in identifying false negatives and other groups of workers that could reasonably be classified as false positives (e.g., "displaced workers" whose jobs were temporary or who had returned to work for their former employers for spells of undetermined length).

Recognizing the implications of these findings, program managers scheduled "forums" with internal and external subject-matter experts that served to clarify various conceptual issues and determine user needs (see Table 1, Cell C_{R41}). They also authorized work to expand the scope of the survey to gather data on both voluntary and involuntary separations, and this decision had implications for the design of the third evaluation in the series. Internal content and design specialists met on a regular basis to review evaluation data and to discuss questionnaire design issues (e.g., concepts, content, question objectives). In 1999, design work began on a new *job-separations supplement* (**JSS**). Many of the items that had been used so successfully in the respondent debriefing questionnaire were incorporated into the draft of the new questionnaire. In late 1999, key parts of the draft supplement were subjected to preliminary cognitive testing (i.e., eleven socio-cognitive interviews; see Table 1, Cell C_{R64}). From the perspective of the framework provided earlier, these redesign activities might best be classified as work within the conceptualization, operationalization and evaluation phases of the questionnaire-redesign-and-evaluation process (Table 1, Phases RP3 through RP6).

The careful reader may have noticed that there was no mention of RP1 activities in the prior paragraph, which refer to the observational work that supports subsequent redesign activities. While there was no specific research subsequent to phase two that involved the

direct and systematic observation of job separations in natural contexts, content specialists at the BLS did have a substantial amount of information available that might qualify as indirect observations. That information came from four sources: (1) evaluation metadata from prior research efforts (phases one and two) – particularly respondent debriefing metadata (e.g., verbatim entries from the "other/specify" precodes) and feedback from interviewers regarding problematic cases; (2) several forums with subject-matter experts; (3) an ongoing review of relevant books and periodical materials (e.g., BLS reports; journal articles; newspaper reports), and (4) communications with BLS staff working on parallel programs.

5.3. The third evaluation: 2000

This third evaluation was organized to accomplish two primary goals: (1) to field-test a set of items that were being considered for the new survey/supplement on job separations (i.e., a subset of the respondent debriefing questions; see Table 9 for examples), and (2) to gather data on two hypotheses that might help to explain why the DWS was not identifying all displaced workers (i.e., some potential false negatives). The vehicle for accomplishing these goals was the respondent debriefing questionnaire. A secondary goal was to continue gathering evaluation data on supplement items SD1 and SD2. So, as part of this administration of the DWS in the year 2000, two forms of evaluation were occurring simultaneously: limited quality assessment work on the DWS and field-based pretesting work on various items under consideration for the expanded survey on job separations – P8 and RP6, respectively.

5.3.1. Methodological details and findings

Relative to prior evaluations, this third evaluation was moderate in terms of size. With respect to interviewer debriefing, two focus groups were conducted with 22 CPS interviewers serving as research participants. The questions used to debrief interviewers on SD1 and SD2 were very similar to those asked in the prior phases. Though we had expected little new from interviewers with respect to SD1 and SD2, such was not the case (see Table 3). Many of the problems identified by interviewers had to do with fundamental conceptual issues: What constitutes a job? When is "leaving a job" indistinguishable in principle from losing a job?

The questions used to debrief interviewers on the prospective items for the new survey on job separations can be found in Table 9 (see bullets under items SDB1 through SDB6). In the JSS, item SDB3 is being considered as the opening screener question in the hope that it will minimize respondent burden. The function of SDB3 would be to identify persons who lost or left a job during a one-year reference period for any reason. The distinction between displacements and job separations that are not displacements would be based on responses to subsequent questions. Some of the problems noted above for SD1 and SD2 are relevant for SDB3 as well. However, other examples of problematic situations surfaced here (see Table 3). Expanding the scope of the supplement to gather data on both voluntary and involuntary separations will place substantial demands on both interviewers and respondents (see Table 9, items SDB2A and SDB2B, for wording). Operationally, this set of questions can pose challenges for interviewers, who first have to

Table 9. Examples of interviewer debriefing questions (bullets) for selected respondent debriefing items (Phase three, 2000)

SDB1	Earlier you told me that you had lost or left a job during the period 1997 through 1999 because (*fill with displacement reason from SD2*). Did you lose that job or did you leave that job? (Note: This question is only asked of respondents who answered "yes" to supplement item SD1 and answered SD2 with precodes 1-3 and 6.) • Did any respondents appear to have difficulty understanding what was meant by the term "job"? • Did any respondents appear to have difficulty understanding the difference between losing a job or leaving a job? • Did you notice any special problems that proxy respondents might have had in answering this question?
SDB2A	People lose jobs for a variety of reasons. In some cases, the person may have experienced problems with a boss or have been let go for poor performance. In other cases, the person's employer may have closed down the company or cut back on jobs. What is the MAIN reason you no longer work for your former employer? (Note: 20 substantive response options were provided, seven referred to employer actions (e.g., employer was: closing down company; moving, merging or selling company; cutting back jobs; downsizing, reorganizing, or outsourcing jobs) and 13 referred to personal reasons/actions (e.g., to take job with better pay; to start own business; problems with boss or employer; problems with old job; own illness/injury; family obligations; to attend school; quit job).) • Did you have difficulty reading this question in its entirety? • Did any respondents have difficulty identifying the MAIN reason why the target person is no longer working for her/his former employer? • Did you experience any difficulty matching respondent's answers to the available response options? If YES: What types of coding problems did you encounter? • Did the list of options (1 – 19) seem to cover most reasons provided by respondents? If NOT: What types of responses did you categorize as "other reasons" (20)?

Table 9. Continued

SDB2B	Some people leave jobs for personal reasons, such as to further their education or to start their own business. Others leave jobs they would have preferred to keep, perhaps because their employer was closing down the company or cutting back on jobs. What is the MAIN reason you no longer work for your former employer? (Note: See SDB2A for examples of response options and debriefing questions.)
SDB3	During the ONE-YEAR period, January through December 1999, did you lose or leave (or retire from) any full or part-time job? (Note: This and following debriefing items are only asked of respondents who answered "no" to supplement item SD1.) • (Note: See first two bullets/questions listed for SDB1.) • Did any respondents appear to have difficulty understanding the difference between a full-time job and a part-time job?
SDB3B	How many jobs, total, did you lose or leave during 1999? • (Note: See first bullet/question listed for SDB1.) • Did any respondents have difficulty recalling how many jobs the person had lost during 1999?
SDB4	(We'd like to focus NOW on the job that was held for the LONGEST TIME:) Did you lose that job or did you leave that job? • (Note: See first two bullets/questions listed for SDB1.)
SDB5A	(Note: Same wording as SDB2A, but asked of respondents who lost a job and answered "no" to SD1.)
SDB5B	(Note: Same wording as SDB2B, but asked of respondents who left a job and answered "no" to SD1.)
SDB6	Was the job (you/she/he) (lost/left/retired from) a TEMPORARY job, that is, a job that was supposed to last only for a limited time or until the completion of a project? • Did any respondents appear to have difficulty understanding the concept of a "temporary job"?

extract the essence of a respondent's sometimes lengthy and/or vague answer, and then find a match for that answer among 20 available response precodes. Several interviewers had difficulty selecting the appropriate precodes, given information provided by respondents. On occasion, when either of two precodes seemed appropriate, the interviewer would read both and have the respondent decide which one sounded best. The problem here, and in any situation where a respondent's initial answer is vague or convoluted, is the potential error that is introduced as a byproduct of interviewers decoding, abstracting, and then matching a respondent's answer to a long list of precodes.

With respect to behavior coding, 60 household interviews were monitored and 131 person-interviews were coded. Much as in the first two phases, interviewers struggled with the wording of both SD1 and SD2, especially the latter, and respondents experienced difficulties providing an adequate answer to SD2 (see Table 5). The screener (SDB3) and reason-for-separation questions (SDB2A/2B and SDB5A/5B) that are being considered for the JSS appeared to outperform their counterparts on the DWS, items SD1 and SD2, respectively.

To this point, when presenting behavior-coding findings, we have limited the presentation to summary statistics. Even though taking notes during live coding can be difficult, it is often possible to gather some qualitative data during the coding process. When this happens, the exchange record can be quite informative (see Table 10).

With respect to respondent debriefing, the debriefing questionnaire comprised 11 items. The wording of items SDB1 through SDB6 can be found in Table 9; items SDB7 and SDB8 are discussed below. Sample sizes for these items ranged from $n = 122$ to $n = 20{,}393$. Using these debriefing data (and other adjustments based on data from several key supplement questions), it was possible to estimate in an approximate fashion the percentage of potential false negatives for the year 1999, and that figure is 29 percent. Similar to phase two results, more than two-thirds of this potential measurement error is associated with presumably inaccurate responses to supplement item SD1. We were curious to know why a respondent would answer "no" to SD1 and, later, answer a series of debriefing questions in such a way as to suggest that the target person was indeed displaced from a job. There would appear to be no shortage of hypotheses regarding this puzzling finding (e.g., overlooking a marginal or a part-time job; data-entry errors; a fatigue effect; a respondent conditioning effect; purposeful misreporting; evaluation-based error). While there may be some truth to all such hypotheses, we choose to focus on two that seemed particularly plausible. The first hypothesis was that some respondents may have overlooked separations from jobs at which the target person worked relatively few hours per week; this hypothesis was tested using debriefing item SDB7 ("How many hours per week did you USUALLY work at that job?"). The second hypothesis was that some respondents may have overlooked separations from secondary jobs for persons who were multiple-job holders; this hypothesis was tested using SDB8 ("At the time you (*fill*: "lost" or "left") that job, were you working at another job?). Data from SDB7 indicate that a majority of persons identified as false negatives (76.4 percent) had worked at jobs that we would characterize as full time (i.e., 36 or more hours per week). Data from SDB8 indicate that a large majority of persons identified as false negatives (90.3 percent) had not been working at another job when they were displaced. While some persons displaced from jobs may have been missed because they worked relatively few hours at their jobs

Table 10. Behavior-coding protocols containing possible classification errors (Phase three, 2000)

Item	Response	Interviewer entry
Case 49	*Protocol (Second person in household)*	
SD1	The respondent answered, yes, that he had *left* a job.	Precode 1 (yes).
SD2	The respondent answered that he was "terminated" from a job with (ABC Oil) (pseudonym).	Interviewer probed to see if precode 3 (position or shift abolished) was acceptable and apparently it was.
Note:	(Intervening questions omitted.)	
SD5	No. (SD5 asks if he had received advance notice of the impending separation.)	Precode 2 (no).
Note:	(Intervening questions omitted.)	
SDB1	Lost job.	Precode 1 (lost job).

Comments (Case 49): When a person says he was "terminated" from a job, it sometimes can mean that he was "fired for cause" (e.g., poor performance), which excludes the person from being counted as a displaced worker. If that is true here, then this case would represent a classification error (i.e., false positive), because it looks like the target person will be classified as a displaced worker based on the way his responses were recorded. However, it is also possible that the respondent provided information – perhaps missed during the coding process – suggesting that his position was being abolished for economic reasons. If that is the case, then there is no classification error. Also worth noting is how the respondent volunteered in SD1 that he had left a job and then, in answering debriefing item SDB1, stated that he lost that job. For some people, to admit losing a job is embarrassing, so they initially respond by saying they left a job. But such self-presentation strategies can have an effect on displacement estimates if persons who leave jobs have to satisfy certain additional conditions (e.g., written advance notice) that persons who lose jobs do not have to satisfy to be classified as displaced.

Table 10. Continued

Item	Response	Interviewer entry
Case 53 Protocol (First and only person in household)		
SD1	Yes, "downsizing".	Precode 1 (yes).
SD2	No response was recorded. Interviewer read first part of SD2 and stopped – she/he did not read any of the reasons and probably just verified that the person had been downsized.	Precode 6 (other).
Note:	(Intervening questions omitted.)	
SDB1	Lost job.	Precode 1 (lost job).
SDB2A	"Cut back" at former place of employment (retail store). The respondent said something like "the department doesn't exist anymore."	Precode 3 (employer was cutting back or eliminating person's job, position or shift). (Precode 5 (other employer actions: downsizing, reorganization, . . .) also would have been acceptable here.)
Comments (Case 53): There is no ambiguity about this case; the person should have been classified as displaced and, based on the way his responses were recorded, that will not happen (i.e., false negative).		

(hypothesis one) or because they held more than one job (hypothesis two), data from debriefing items SDB7 and SDB8 do not provide strong support for either hypothesis.

5.3.2. Discussion and implications for future research

In addition to corroborating prior findings, this third evaluation provided useful information/data regarding an alternative set of questions for identifying displaced workers and counting job separations. These new questions appear to outperform SD1 and SD2 in certain respects. For example, on the basis of behavior coding data, debriefing item SDB3 is certainly an easier screener question to read and to answer than supplement item SD1 (see Table 5). The new reason-for-separation items (e.g., SDB2A and SDB2B) with their free-response, field-coded format are easier for interviewers to read as worded and yield a fairly high percentage of adequate answers relative to SD2; moreover, these new items appear to be successful at capturing displaced workers that SD1 and SD2 miss. Such findings are encouraging; however, for a variety of reasons, one should resist the temptation to conclude that these new items will necessarily produce a more accurate estimate of displaced workers than the current supplement. First, the efficacy of the respondent debriefing items is not independent of the supplement items they were designed to evaluate. To illustrate, items SDB2A and SDB2B essentially reassess cases that SD2 initially rejected as not belonging within the displaced-worker category; these debriefing items did not have to shoulder the full burden of identifying displaced workers independently. Second, data from behavior coding and interviewer debriefing suggest that these new items are not immune to some of the same conceptual problems (e.g., what counts as a job; how is losing a job different from leaving a job) and operational problems (e.g., difficulty matching responses to specific precodes) that bedevil SD1 and SD2. And third, we have yet to validate the data generated by either set of displacement questions. If we are to have confidence in the utility and validity of these new items, they will have to be evaluated independently.

6. Discussion

In the introduction to this article, four ways in which this research might contribute to questionnaire-evaluation practice and theory were noted. Those aspirations are revisited in this closing section.

6.1. Benefits of iterative, multiple-method questionnaire evaluation research

Given scarce resources and a mandate to assure high-quality data, survey sponsors and program managers have difficult decisions to make regarding the collection of evaluation metadata (Hert 2002). How much funding and staff time can be allocated to questionnaire evaluation research? When and how frequently should such research be conducted? How extensive should the research be? While not intending to tax the limited resources of survey sponsors, it is possible to enumerate some of the benefits of conducting iterative, multiple-method questionnaire evaluation research (hereafter, simply iterative research) based on experiences described herein.

One of the benefits of iterative research is its confirmation potential. Replications, even partial replications, inform researchers as to what findings can be trusted and

which findings inspire less confidence. For example, we have considerable evidence to suggest that there is measurement error associated with SD2, especially with respect to precode 6 ("some other reason"). We can quantify this error in a crude sort of way and offer plausible explanations for its existence (e.g., uncertainty regarding how to code certain responses; inadequate question specifications). In contrast, though we have consistent quantitative evidence to suggest that displaced workers are being missed as a result of inaccurate "no" responses to SD1, we are unable to offer a more compelling explanation for these false negatives than poor question design. As a result, the latter finding inspires less confidence. A second benefit of iterative research is its educational value with respect to methodology. With each successive evaluation phase, one comes to appreciate each method's special character, its potential and its limitations, its similarities and differences with respect to other methods and techniques (cf. Groves 1996; Presser and Blair 1994; Rothgeb, Willis, and Forsyth 2001) – more on this topic later. We learn, too, about our own limitations and fortunately have the opportunity to consider methodological improvements in subsequent phases. For example, by modifying the content and increasing the number of respondent debriefing questions asked in phase two, we not only once again found evidence that we were missing/misclassifying some persons who should have been counted as displaced workers (false negatives), but also that we may have been counting some workers as displaced who really should not have been so classified (false positives; e.g., temporary workers). A third benefit of iterative research is its educational value with respect to question/questionnaire design. Through repeated observations, design specialists who conduct iterative research learn which types of questions do not work; in the process, we also learn how to craft better questions. For example, it became obvious after phase two that SD1 is a poor screener question. It imposes unnecessary burden on respondents, most of whom have not been displaced from a job, and it invites erroneous answers with the ambiguous phrase, "or another similar reason." With regard to design, the verbatim entries from respondent debriefing questions asked in early evaluation phases were helpful in developing response precodes for debriefing questions used in later phases, some of which are now being considered for use in the new survey on job separations.

Some readers will be concerned, understandably, about the costs associated with iterative questionnaire evaluation research. On that concern, consider two points. First, costs can usually be controlled by limiting the scope of research and, in the case of governmental sponsorship, by assigning work to in-house staff. Much of the evaluation work reported above was accomplished by one survey methodologist – in collaboration with content specialists and operations and field staff. Second, periodic evaluations/monitoring may be one of the more efficient strategies for tracking and adjusting to substantive change in rapidly evolving subject-matter domains.

6.2. The utility of a multiple-method approach to evaluating questionnaires

The three evaluation methods used in the present research effort attempt to capture or reveal the perspectives of various informational sources. Interviewer debriefings capture the perspectives of interviewers and, in an indirect and filtered way, reveal some of the

difficulties experienced by respondents. Respondent debriefings capture the perspectives of survey-eligible individuals (and their proxies), but only with respect to the specific interests and goals of content and design specialists, whose perspectives are also revealed as part of the process. Behavior coding, a relatively unobtrusive and objective method, captures the essence of the question-and-answer process and in so doing reveals the observable difficulties interviewers and respondents may be experiencing within a particular context. While a multiple-method evaluation strategy provides no guarantee that all significant antecedents of measurement error will be detected, it does place the research team in a good position to identify specific antecedents (e.g., poor question design; confusing or inadequate item specifications; inappropriate probing). To the extent that a particular evaluation strategy is successful at identifying the most significant antecedents of measurement error, the strategy can be said to possess diagnostic utility. To the extent that such findings are helpful in making informed decisions regarding the development of a new questionnaire or the redesign of an existing one, the strategy can be said to possess *design utility*. The two forms of utility are not necessarily highly correlated (e.g., Forsyth, Rothgeb, and Willis 2004).

Adopting an economic metaphor, one could also consider the productivity associated with specific evaluation techniques (i.e., "information yield" divided by "labor investment"). In a very general and subjective sense, information yield would refer to the amount of useful information/data generated by a particular technique and labor investment would refer to the amount of effort expended by the research team in implementing the technique and in analyzing information/data (cf. Groves 1996; Presser and Blair 1994; Rothgeb, Willis, and Forsyth 2001). The yield associated with a particular technique would depend, in part, on the manner in which it is applied by the research team in a given context. Specific applications of a particular technique vary greatly in terms of how much human (versus machine) effort is involved in collecting, analyzing and summarizing information/data. For example, a survey methodologist could choose to debrief interviewers by conducting one focus group, or say five. The focus group could be designed to run for one hour or two. The moderator could rely on notes taken during the session, or could transcribe and summarize information captured on audiotape. Increasing the labor component may increase the yield of a particular technique, but if yield does not increase more than proportionally, productivity may actually decline or remain stable.

With respect to this effort – specifically phase-two research – a subjective impression of the productivity associated with various evaluation techniques is presented in Table 11. The follow-up probe technique had the highest productivity score (P) and the largest values for both information yield (Y) and labor investment (L). The fact that we could use respondent debriefing data to estimate measurement error and to address certain conceptual/specification issues in a quantitative manner (e.g., the percentage of persons who worked at temporary jobs; the percentage of persons who lost jobs as opposed to leaving them) made this a very useful and powerful evaluation technique. Moreover, respondent debriefing data have enormous surplus value in that any number of potentially informative cross-tabulations can be run with other debriefing items, or with supplement items, as the need arises. The focus groups were also quite productive, primarily with respect to identifying conceptual problems. Behavior coding was useful in quantifying problems with the question-and-answer process and in corroborating findings from other

Table 11. A subjective assessment of productivity, information yield and labor investment associated with four questionnaire evaluation techniques (Phase two, 1998)

Method/Technique	*P*	*Y*	*L*	Comments
Interviewer debriefing				
Focus group	1.25	5	4	• Qualitative data: Retrospective and subject to situational effects (e.g., group dynamics). • Useful for identifying conceptual and operational problems. • Sample of interviewers not representative of population. • Provides no quantitative basis for estimating measurement error.
Rating form	1.00	1	1	• Descriptive quantitative ratings data: Retrospective and potentially contaminated if interviewers talk about items. • Useful in identifying differences among interviewers, but sample of interviewers not representative of population. • Minimal labor on part of researcher. • Provides no quantitative basis for estimating measurement error.
Interaction coding				
Behavior coding (live)	1.00	3	3	• Descriptive quantitative data and some qualitative data. • Useful in detecting possible problems with specific items, but not necessarily useful in identifying solutions. • Useful for comparative analyses (open vs. closed questions) • Relatively objective/unbiased. • Sample of interviewers and respondents not fully representative of their respective populations. • Live coding more susceptible to error and omissions than other coding strategies (e.g., coding from audiotapes). • Provides no quantitative basis for estimating measurement error.

Table 11. Continued

Method/Technique	*P*	*Y*	*L*	Comments
Respondent debriefing				
Response-dependent follow-up probes	1.50	9	6	• Quantitative data: Useful in confirming/quantifying specification problems (see last bullet). • Expandable, as need arises, as cross-tabulations can be run with other debriefing items and with items from the host questionnaire. • Qualitative data: "Other-specify" precodes provide quasi-ethnographic data. • Respondent sample fairly representative of population. • Adds to respondent burden in some cases. • Labor intensive for content and design specialists. • Potentially very useful in estimating measurement error associated with specific items. However, potentially misleading if questions are not balanced with respect to identifying false positives and false negatives.

Note: Productivity (*P*) equals yield (*Y*) divided by labor (*L*). Values for *Y* and *L* are based on a subjective ten-point rating scale with ordinal scale characteristics.

techniques. It is important to recognize that each of these techniques was designed with a specific goal in mind. Had the context been different (e.g., target questionnaire, available resources, experience level with respect to methods), the scores and values associated with these techniques would have been different as well.

As many practitioners have noted, each of these techniques possesses certain weaknesses. With regard to the use of follow-up probes, it is not always clear what probe questions one might need to ask and, even when an objective for a probe is clear, one may not be completely successful in achieving that objective. For example, in phase two, a debriefing question was asked to determine if the job a person lost or left for a displacement reason was a temporary job: "Was the job you lost a temporary job, that is, a job that was supposed to last only for a limited time or until the completion of a project?" The expectation was that a large majority of persons for whom a "yes" answer was provided would have worked at such jobs for relatively brief periods of time (e.g., six months or less). When the debriefing item was cross-tabulated with a supplement item on employment duration ($n = 108$), it was found that approximately 40 percent of displaced workers had worked for their employer for more than a year and that 25 percent had worked for more than two years. In other words, probe questions can be just as problematic as the questionnaire items they are designed to evaluate. With regard to interviewing debriefing techniques, focus groups are highly susceptible to group dynamics and, depending on how research participants are selected, may not be representative of the interviewer population. Retrospective rating forms are subject to memory or salience effects, and occasionally yield findings that are difficult to explain. For example, in phase two, interviewers at two telephone centers had identified numerous problems with SD2; when asked to rate this item, 12 of 22 interviewers gave it relatively high difficulty ratings (3-to-5 range). Quite inexplicably, not one interviewer in a group of twelve at the third telephone center identified SD2 as problematic (see Table 7) and, as a result, SD2 was not rated at that location. With regard to behavior coding, the principal weakness associated with this technique – given the manner in which we chose to employ it – is that, while it is useful in identifying where problems exist, it provides little guidance as to what may be causing these problems. Another weakness – associated more with the coder (the present author) than with the technique – was that only interviews with English-speaking respondents could be monitored and coded.

As the discussion above suggests, all methods and techniques used to evaluate questionnaires have inherent weaknesses. Relying on any one method or technique is risky. The adoption, then, of a multiple-method evaluation strategy serves two purposes: (1) it minimizes the risk associated with single-method evaluations, and (2) it captures the perspectives of the various interdependent sources that contribute to measurement error. Rather than being viewed as a means for discovering "truth," a multiple-method evaluation strategy is more about developing an understanding (via triangulation) of what might be problematic regarding a particular questionnaire item or set of items. It is this understanding that enables content and design specialists to pursue remedial action (e.g., informed design modifications; full-scale redesign).

6.3. The importance of clear and well-grounded conceptual specifications in minimizing measurement error

Inadequacies in conceptual specification, be they manifested in question wording or in questionnaire-relevant metadata (e.g., question objectives; interviewer instructions), greatly complicate design and evaluation processes and the assessment of both response and measurement error (Federal Committee on Statistical Methodology 1988; Hox 1997; Martin 1987; Turner and Martin 1984, Chapter 7). Freedman (Federal Committee on Statistical Methodology 1988, p. 34) refers to such inadequacies as specification error, which he defines as "the error that occurs at the planning stage of a survey because data specification is inadequate and/or inconsistent with respect to the objectives of the survey." He states further that: "Specification error can result simply from poorly worded questionnaires and survey instructions or may reflect the difficulty of measuring abstract concepts." For example, how could it be that the following verbatim responses (SD2, precode 6) were not coded/classified as displacement reasons (see Esposito 2002, Appendix, for relevant DWS metadata): "company merged with another company," "laid off permanently," "employer cut person's hours," "office closed and had to move," "because of the Asian stock market crash," "pushed out of position," "program was not refunded," "company couldn't afford her services anymore," "business was sold," "never called back to work," "company was part of acquisition by other company." The miscoding of these responses represents avoidable measurement error. Not so easy to pinpoint, however, is the source or sources of that error: question wording, conceptual specifications, interviewer training – all of the above.

The objective of this discussion is not to disparage the DWS, its sponsor or its designers: Designing effective and cost-efficient questionnaires is difficult work and, as noted, the DWS was never intended as a panel survey. The objective is to drive home two points. First, well-crafted conceptual specifications, grounded in current, domain-relevant observations, are critical if we are to succeed in minimizing measurement error. And second, we must accept the fact that all of the important social domains we seek to measure are changing, evolving – some faster, some slower. Unless we monitor this change, note when disparities become significant and make the appropriate modifications to our survey instruments and associated metadata, measurement error will gradually undermine the quality and the utility of the data and information we disseminate.

6.4. The potential utility of a broad organizational framework in addressing and solving problems of a theoretical and applied nature

The framework draws attention to several important issues: (1) the inextricable relationship between questionnaire *design* and questionnaire *evaluation* processes; (2) the complex and variable interrelationships among various sources of measurement error across potentially recursive design-and-evaluation phases; and (3) the inevitability and relativity of change in various target domains.

With respect to the first item, we need to be more explicit in describing the types of activities that take place during the early design-and-evaluation phases. For example, some of us may equate operationalization with the not-so-simple task of converting a survey sponsor's question objectives into a reasonable set of questionnaire items.

The metadata literature suggests that there is much more to this process than drafting individual questionnaire items. What are the crucial aspects of the operationalization process? What methods do we have for evaluating the various aspects of this process? When requisite tasks have been performed poorly (e.g., vague definitions; inadequate training materials), how is measurement error affected? What happens (or does not happen) during latter design-and-evaluation phases is also important. For example, what are the implications of conducting perfunctory evaluations, either pretesting (P6) or quality assessment (P8), or of dismissing such evaluations altogether? What would be the implications of not involving design specialists in the operationalization phase (Cell C_{52}) or of not involving content specialists in the evaluation process (Cell C_{81})?

With respect to the second item above, we would encourage researchers to explore various relationships between and among the cells in Table 1. For example, we accept as axiomatic that activities associated with superordinate cells in a particular column have the potential to influence activities associated with subordinate cells (and vice versa), as long as the phases involved are recursive (i.e., recycling between Phases P1 through P6). Consider Column one: Being involved in pretesting work (Cell C_{61}) could cause content specialists to revisit C_{11} and C_{31} activities (observation and conceptualization, respectively), because recursive movement is possible between P1 and P6. In principle, it is not possible for activities at C_{81} (quality assessment work) to affect activities at C_{11}, because survey administration (P7) acts as a barrier to recursivity. However, activities at C_{81} may have an effect on subsequent observational activities of a content specialist (C_{R11}), even though no formal redesign effort may yet be underway (see below). We might also consider the interrelationships among various sources of measurement error that exist during any one phase of the questionnaire design-and-evaluation process (e.g., P7). A review of the behavior coding protocols in Table 10 provides some insights into the nature of these interrelationships. For example, had content specialists updated DWS instructions to provide guidance to interviewers regarding downsizing (e.g., code as "position abolished"), Case 53 probably would not have resulted in a false negative. Had the interviewer actually read Item SD2, perhaps the respondent would have selected one of the three displacement reasons. Had design specialists modified the wording of SD1 and SD2 at some point after 1984 to reflect changes in the economy, perhaps there would have been a different outcome. Had the interview been conducted face-to-face rather than by telephone, perhaps the interviewer or the respondent would have behaved differently. Such protocols illustrate what we have been alluding to as the collaborative nature of measurement error (and accuracy).

With respect to the third item above, and given that change is inevitable and relative, how does one recognize significant/substantive change in a target domain – that is, some fundamental change that alters meaning and in so doing threatens the integrity of a statistical time series? Consider an example: Rather than use the term "layoff" to describe staff reductions, some human resource professionals prefer the terms "downsizing" or "restructuring," which are relatively recent euphemisms. At present, these terms are not specifically identified as displacement reasons in the DWS or its instructional metadata and, as a result, the use of them may increase measurement error to some extent. Do such changes in terminology represent superficial change or substantive change? If the former, one might then ask: Do ongoing and relatively minor modifications to a questionnaire

(and associated metadata), designed to minimize the negative effects of superficial content changes in a target domain, constitute a threat to the integrity of a time series? These are not questions with easy answers. Content and design specialists need a vocabulary to describe and track domain-specific change, and a set of principles or standards to help them decide when such changes require that remedial action be taken (e.g., adjustments in questionnaire content and metadata; redesign activities).

Minimizing the measurement error associated with our most important social surveys needs to be viewed as an ongoing process. That process starts with observation, and regular monitoring and periodic evaluations of policy-relevant, panel surveys should be viewed as an essential component of the process, not as an option. Every aspect of this process is important – so too every participant and every specialized group. If we are to succeed in our efforts to produce high-quality statistical data, constructive collaboration will be essential.

7. References

Abraham, K.A. (1996). Ensuring Quality in the Data Collection Process. Commissioner's Order No. 2-96. Washington, DC: U.S. Bureau of Labor Statistics.

Akkerboom, H. and Dehue, F. (1997). The Dutch Model of Data Collection Development for Official Surveys. International Journal of Public Opinion Research, 9, 126–145.

Barsalou, L.W. (1992). Cognitive Psychology: An Overview for Cognitive Scientists. Hillsdale, NJ: Lawrence Erlbaum.

Beatty, P. (1995). Understanding the Standardized/Non-Standardized Interviewing Controversy. Journal of Official Statistics, 11, 147–160.

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

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.

Campanelli, P.C., Martin, E.A., and Creighton, K.P. (1989). Respondents' Understanding of Labor Force Concepts: Insights from Debriefing Studies. Proceedings of the Fifth Annual Research Conference. Washington, DC: U.S. Bureau of the Census, 361–374.

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

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

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

Conrad, F.G. and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly, 64, 1–28.

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

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

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

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

Dippo, C. and Sundgren, B. (2000). The Role of Metadata in Statistics. Proceedings of the American Statistical Association, Second International Conference on Establishment Surveys: Survey Methods for Businesses, Farms and Institutions. Alexandria, VA: American Statistical Association, 909–918.

Esposito, J.L. (2002). Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study. Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing (QDET) Methods, Charleston, SC.

Esposito, J.L. (2003). A Framework Relating Questionnaire Design and Evaluation Processes to Sources of Measurement Error. Proceedings of the 2003 Federal Committee on Statistical Methodology Research Conference. Statistical Policy Working Paper 37. Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Esposito, J.L. and Fisher, S. (1998). A Summary of Quality Assessment Research Conducted on the 1996 Displaced-Worker/Job-Tenure/Occupational-Mobility Supplement. Bureau of Labor Statistics' Statistical Note Series, Number 43. Washington, DC: U.S. Bureau of Labor Statistics.

Esposito, J.L. and Rothgeb, J.M. (1997). Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 541–571.

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 Annual Meeting of the American Association for Public Opinion Research, Danvers, MA.

Federal Committee on Statistical Methodology (1988). Measurement of Quality in Establishment Surveys. Statistical Policy Working Paper 15. Washington, DC: Statistical Policy Office, U.S. Office of Management and Budget, 33–42.

Flaim, P.O. and Sehgal, E. (1985). Displaced Workers of 1979–83: How Well Have They Fared? Monthly Labor Review, 108(6), 3–16.

Foddy, W. (1993). Constructing Questions for Interviews and Questionnaires. Cambridge, UK: Cambridge University Press.

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

Forsyth, B.H., Rothgeb, J.M., and Willis, G.B. (2004). Does Pretesting Make a Difference? An Experimental Test. In Methods for Testing and Evaluating Survey Questionnaires. S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: Wiley Interscience, 525–546.

Fowler, F.J. and Mangione, T.W. (1990). Standardized Survey Interviewing. Thousand Oaks, CA: Sage.

Fowler, F.J. (1995). Improving Survey Questions: Design and Evaluation. Thousand Oaks, CA: Sage.

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

Fowler, F.J. and Cannell, C.F. (1996). Using Behavior Coding to Identify Cognitive Problems with Survey Questions. In Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass, 15–36.

Gerber, E. (1999). The View from Anthropology: Ethnography and the Cognitive Interview. In Cognition and Survey Research, M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds). New York: Wiley, 217–234.

Glaser, B.G. and Strauss, A.L. (1999). The Discovery of Grounded Theory. New York: Aldine De Gruyter.

Goldenberg, K.L., Anderson, A.E., Willimack, D.K., Freedman, S.R., Rutchik, R.H., and Moy, L.M. (2002). Experiences Implementing Establishment Survey Questionnaire Development and Testing at Selected U.S. Government Agencies. Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing (QDET) Methods, Charleston, SC.

Groves, R.M. (1996). How Do We Know What We Think They Think Is Really What They Think? In Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass, 389–402.

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

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

Hert, C.A. (2002). Creation and Use of Metadata in Two Bureau of Labor Statistics Survey Efforts: An Ethnographic Investigation of a Community of Practice. Joint UNECE/ Eurostat Work Session on Statistical Metadata. Working Paper Number 15. Eurostat, Luxembourg.

Hess, J.C. and Singer, E. (1995). The Role of Respondent Debriefing Questions in Questionnaire Development. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA, 1075–1080.

Hox, J.J. (1997). From Theoretical Concept to Survey Question. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 47–69.

Jobe, J.B. and Herrmann, D. (1996). Implications of Models of Survey Cognition for Memory Theory. In Basic and Applied Memory Research (Volume 2): Practical Applications. D. Herrmann, M. Johnson, C. McEvoy, C. Herzog, and P. Hertel (eds). Hillsdale, NJ: Erlbaum, 193–205.

Jobe, J.B. and Mingay, D.J. (1989). Cognitive Research Improves Questionnaires. American Journal of Public Health, 79, 1053–1055.

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213–236.

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

Maynard, D.W. and Schaeffer, N.C. (2002). Standardization and Its Discontents. In Standardization and Tacit Knowledge. D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 3–45.

Miller, K. (2002). Cognitive Analysis of Sexual Identity, Attraction and Behavior Questions. Cognitive Methods Staff, Working Paper Series, No. 32. Hyattsville, MD, USA: National Center for Health Statistics.

Morton-Williams, J. (1979). The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire. Quality and Quantity, 13, 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, 109–127.

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

O'Muircheartaigh, C. (1999). CASM: Successes, Failures, and Potential. In Cognition and Survey Research, M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds). New York: Wiley, 39–62.

Platek, R. (1985). Some Important Issues in Questionnaire Development. Journal of Official Statistics, 1, 119–136.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? In Sociological Methodology, Volume 24, P.V. Marsden (ed.). Washington, DC: American Sociological Association, 73–104.

Rothgeb, J., Willis, G., and Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Montreal, Canada.

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

Schaeffer, N.C. and Dykema, J.L. (2004). A Multiple-Method Approach to Improving the Clarity of Closely Related Concepts: Distinguishing Legal and Physical Custody of Children. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J.M Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: Wiley Interscience, 475–502.

Schuman, H. (1966). The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review, 31, 218–222.

Schwarz, N. and Sudman, S. (eds). (1996). Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

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

Sirken, M.G., D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds). (1999). Cognition and Survey Research. New York: Wiley.

Smith, E.E. (1989). Concepts and Induction. In Foundations of Cognitive Science, M.I. Posner (ed.). Cambridge: MIT Press, 501–526.

Snijkers, G. (2002). Cognitive Laboratory Experience: On Presenting Computerized Questionnaires and Data Quality. Ph.D. Dissertation. Utrecht University, Utrecht, and Statistics Netherlands, Heerlen.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-Face Survey Interviews. Journal of the American Statistical Association, 85, 232–253.

Sudman, S. and Bradburn, N.M. (1974). Response Effects in Surveys. Chicago: Aldine.

Sudman, S. and Bradburn, N.M. (1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass.

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

Tourangeau, R. (1984). Cognitive Science and Cognitive Methods. In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, T. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds). Washington, DC: National Academy Press, 73–100.

Tourangeau, R., Rips, R.J., and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge, UK: Cambridge University Press.

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

U.S. Bureau of the Census (1998). Pretesting Policy and Options: Demographic Surveys at the Census Bureau. Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census (2000). CPS Field Representative Memorandum Number 2000-02 (Field Division). Washington, DC: U.S. Department of Commerce.

Webb, E.J., Campbell, D.T., Schwartz, R.R., and Sechrest, L. (1966). Unobtrusive Measures: Non-reactive Research in the Social Sciences. Chicago: Rand McNally.

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.

Received February 2003
Revised February 2004