

## **ITERATIVE, MULTIPLE-METHOD QUESTIONNAIRE EVALUATION RESEARCH: A CASE STUDY**<sup>1</sup> [For suggested citation, see page 36 of this paper.]

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

*Key Words:* Pretesting, quality assessment, behavior coding, respondent debriefing, focus groups, displaced workers.

### **1. INTRODUCTION**

This paper summarizes a series of three biennial evaluations of a labor force questionnaire that collects data on job 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 (Esposito and Rothgeb, 1997, pp. 543-546). 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.<sup>2</sup> It should be noted at the outset that this series of studies was not iterative by design—it evolved as such.

It is hoped that this paper 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; and (4) by providing a broad organizational framework with which to address and solve problems of both a theoretical and applied nature. In pursuit of these objectives, an organizational framework will be provided that interrelates various phases of the questionnaire design-and-evaluation process with various components of a widely cited model of measurement error. This task is addressed in the following section.

### **2. QUESTIONNAIRE DESIGN, EVALUATION AND MEASUREMENT ERROR: AN ORGANIZATIONAL FRAMEWORK**

The organizational framework borrows and combines elements from two sources. The first source is a rudimentary process model that describes in a very general and simplistic manner

---

<sup>1</sup> *The views expressed in this paper are those of the author and do not reflect the policies of the Bureau of Labor Statistics. Acknowledgements:* This research could not have been accomplished without the contributions and insights of content and design specialists at the Bureau of Labor Statistics (Thomas Nardone, Jennifer Gardner, Jay Meisenheimer, Steven Hipple, Fran Horvath, Jennifer Martel, Alan Eck, and Sylvia Fisher) and without the hard work and cooperation of the Census Bureau's operations and field staff. Thanks also to Nora Cate Schaeffer and Ed Robison for providing very helpful comments to an earlier draft of this paper.

<sup>2</sup> 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 data (i.e., Commissioner's Order No. 2-96, "Ensuring Quality in the Data Collection Process"). This research is also consistent with the Bureau's Quality Measurement Model (1994) and with the pretesting policy of the U.S. Bureau of the Census (1998).

how questionnaires are developed and evaluated (Esposito and Rothgeb, 1997, pp. 543-546, 561-566).<sup>3</sup> The second source is a descriptive model of measurement error that has been articulated by Groves (1987, 1989). The resulting 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

The process model comprises six partially overlapping phases both for initial design and redesign efforts (see **Table 1**):

- **Phase One (P1): *Observation*.** During this initial phase, content specialists focus on observable activity (i.e., behavior and events) within various contexts (family; community; workplace). While the ideal 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.
- **Phase Two (P2): *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. Presumably, content specialists will differ with respect to which concepts they identify as central and with respect to the delineation of causal interrelationships among concepts. An important consideration here, and during the prior phase, is who assumes primary responsibility for observation and 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).
- **Phase Three (P3): *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., interviewing manuals; classification algorithms).<sup>4</sup> At present, this translation/design process would appear to draw as much from art as from science; as a result, one might presume that the experience and expertise of the questionnaire design team would have a significant impact on the magnitude of measurement error.
- **Phase Four (P4): *Evaluation (pretesting)*.** During this phase, design specialists—ideally in close collaboration with content specialists and field operations staff—assume primary responsibility for developing a plan to test the draft questionnaire. This testing usually starts

---

<sup>3</sup> For more thorough discussions of these topics, the reader has many good choices: Akkerboom and Dehue, 1997; Converse and Presser, 1986; Czaja and Blair, 1996; DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Hox, 1997; Forsyth and Lessler, 1991; Fowler, 1995; Oksenberg, Cannell and Kalton, 1991; Platek, 1985; Sudman and Bradburn, 1982; and Turner and Martin, 1984. Hox (1997) provides an eloquent and scholarly discussion on the topics of conceptualization and operationalization. While the approaches he describes for formulating survey questions are very 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 much more empirical and pragmatic.

<sup>4</sup> According to Dippo and Sungren (2000): “The first and most fundamental purpose of metadata is to help the user of statistical data to interpret, understand, and analyze statistical data (microdata, macrodata, or other statistical metadata)... In other words, statistical metadata should help the human user to transform statistical data into information (p. 910).” For other thoughtful discussions on this topic, see Dippo, Conrad and Gillman (2000) and Hert (2002).

with an assessment of how research participants “process” questionnaire content cognitively (i.e., comprehension, retrieval, judgment, response/reporting; Tourangeau, 1984). 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). As noted, a variety of methods have been proposed/described for use in pretesting questionnaires (see citations in footnote 3). When an acceptable draft questionnaire has been developed, it is to the sponsor’s advantage to have the survey instrument (i.e., the questionnaire and its operational components) evaluated in a field setting (e.g., Tucker et al., 1997). If the instrument contains serious design and/or operations flaws, it is far better to find them prior to implementation in a production environment than after.

- Phase Five (**P5**): *Administration*. After pretesting work is completed and modifications are made to the questionnaire and to its pertinent metadata (e.g., interviewer instructions; algorithms), the survey instrument is finalized and moved to a production environment.
- Phase Six (**P6**): *Evaluation (quality assessment)*: Depending on available resources and the importance of a survey’s data products (e.g., monthly unemployment rate; prevalence of heart disease; annual crime rate), 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 (Esposito and Rothgeb, 1997, pp. 543-551).

While socio-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 (P6) in their redesign work (RP1 through RP4). Two notable differences between initial design and redesign phases is the wealth of content-specific and evaluative metadata available during redesign efforts and the potential involvement of questionnaire design specialists during observation (RP1) and conceptualization (RP2) phases.

## 2.2 Interdependent 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 *content specialists* (e.g., subject-matter experts; survey managers or their agents) and *design specialists* (e.g., professionals who: design and evaluate questionnaires, prepare training materials, and develop algorithms). The rationale for this distinction is rooted in the different roles each group assumes in the questionnaire design-and-evaluation process and with the specialized expertise each possesses with regard to resolving certain types of issues and problems (e.g., theoretical/conceptual versus technical design). Brief descriptions of the five sources of measurement error are provided below.

- *Questionnaire (content specialist)*. Especially during the observation and conceptualization phases associated with survey design, content specialists assume a crucial role in describing

the domain of interest, isolating and defining key concepts, and delineating possible relationships among theoretical variables (Federal Committee on Statistical Methodology, 1988; Hox, 1997; Martin, 1987; 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 impact on questionnaire content and data quality. The more accurate their observations and assumptions and theories, the more successful the survey measurement process is likely to be.

- *Questionnaire (design specialist)*. During initial design, the questionnaire-design specialist, usually following guidelines prescribed by researchers and practitioners (Belson, 1981; Converse and Presser, 1986; Foddy, 1993; Fowler, 1995; Sudman and Bradburn, 1982) transforms conceptual specifications provided by the content specialist into coherent sets of questionnaire items and ancillary metadata (e.g., interviewer instructions; classification algorithms). Even when conceptual specifications appear reasonably clear and precise, this translation/design process can be challenging. Compromises that increase the potential for measurement error (e.g., keeping items and questionnaires short) may need to be made to reduce reporting burden and conserve resources. During redesign efforts, design specialists may be enlisted to provide support during observation (RP1) and conceptualization phases (RP2), especially if they participated in prior evaluative research (P6).
- *Interviewer*. With respect to minimizing measurement error, there would appear to be divergent 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). Neither position completely removes interviewers as a potential source of measurement error, and the relative cost (time and money) of error reduction is an important consideration. A reasonable compromise—one that interviewers seem to accept quite naturally—might be to employ standardization as a starting point in questionnaire design and administration, but also to educate and empower interviewers to deviate from standardized scripts in those situations where they believe the communication of intended meaning may be problematic.
- *Respondent*. In an effort to improve data quality, behavioral scientists: have developed socio-cognitive models of the response process (e.g., Cannell, Miller and Oksenberg, 1981; Tourangeau, 1984); have described the types of cognitive errors that can occur at each stage (Tourangeau, Rips and Rasinski, 2000); and have devised various strategies for identifying questionnaire problems and reducing measurement error (Schwarz and Sudman, 1996; Gerber, 1999). Significant 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 have a significant motivational component (e.g., competing demands on the respondent’s time; uninteresting survey content). When unmotivated to participate in a survey, respondents may engage in satisficing behavior (Krosnick, 1991), increasing the likelihood and magnitude of measurement error.

**Table 1. A Framework Relating Questionnaire Design-and-Evaluation Processes to Sources of Measurement Error**

|  |            | <b>Interdependent Sources of Measurement Error (at P5 or RP5) <sup>1</sup></b> |  |   |                        |                        |                        |
|--|------------|--|--|---|------------------------|------------------------|------------------------|
|  |            | <b>Initial Design</b>  | <i>Questionnaire</i><br>[Content Specialist] | <i>Questionnaire</i><br>[Design Specialist] | <i>Interviewer</i>     | <i>Respondent</i>      | <i>Mode</i>            |
| <b>Questionnaire Design and Evaluation Phases <sup>3</sup></b> | P1         | <i>Observation</i>   | C <sub>11</sub>                              |   |                        |                        |                        |
|  | P2         | <i>Conceptualization</i>   | C <sub>21</sub>                              |   |                        |                        |                        |
|  | P3         | <i>Operationalization</i>  | C <sub>31</sub>                              | C <sub>32</sub>                             |                        |                        |                        |
|  | P4         | <i>Evaluation (PT) <sup>4</sup></i>  | C <sub>41</sub>                              | C <sub>42</sub>                             | C <sub>43</sub>        | C <sub>44</sub>        | C <sub>45</sub>        |
|  | <b>P5</b>  | <b><i>Administration</i></b>   | <b>C<sub>51</sub></b>                        | <b>C<sub>52</sub></b>                       | <b>C<sub>53</sub></b>  | <b>C<sub>54</sub></b>  | <b>C<sub>55</sub></b>  |
|  | P6         | <i>Evaluation (QA) <sup>4</sup></i>  | C <sub>61</sub>                              | C <sub>62</sub>                             | C <sub>63</sub>        | C <sub>64</sub>        | C <sub>65</sub>        |
|  |            | <b>Redesign</b>  |  |   |                        |                        |                        |
| <b>Questionnaire Redesign and Evaluation Phases</b>            | RP1        | <i>Observation</i>   | C <sub>R11</sub>                             | (C <sub>R12</sub> ) <sup>2</sup>            |                        |                        |                        |
|  | RP2        | <i>Conceptualization</i>   | C <sub>R21</sub>                             | (C <sub>R22</sub> )                         |                        |                        |                        |
|  | RP3        | <i>Operationalization</i>  | C <sub>R31</sub>                             | C <sub>R32</sub>                            |                        |                        |                        |
|  | RP4        | <i>Evaluation (PT)</i>   | C <sub>R41</sub>                             | C <sub>R42</sub>                            | C <sub>R43</sub>       | C <sub>R44</sub>       | C <sub>R45</sub>       |
|  | <b>RP5</b> | <b><i>Administration</i></b>   | <b>C<sub>R51</sub></b>                       | <b>C<sub>R52</sub></b>                      | <b>C<sub>R53</sub></b> | <b>C<sub>R54</sub></b> | <b>C<sub>R55</sub></b> |
|  | RP6        | <i>Evaluation (QA)</i>   | C <sub>R61</sub>                             | C <sub>R62</sub>                            | C <sub>R63</sub>       | C <sub>R64</sub>       | C <sub>R65</sub>       |

**\* Notes:** (1) The phrase “*interdependent sources of measurement error*” has been adopted to reflect the view that measurement error—and accuracy—are presumed to be the outcome of collaborative/interactive processes. Within a specific context, measurement error is presumed to be a byproduct of role- and task-specific activities that take place during the survey administrative phase, P5 or RP5. Activities that occur at prior phases (e.g., P1 through P3) can be viewed as precursors to measurement error. [Rapid change within a specific target domain over time also leads to measurement error.] (2) Each labeled cell represents role- or task-specific activities (Sudman and Bradburn, 1974) that potentially affect the magnitude of error at a particular phase (and beyond). Parentheses suggest that involvement by a design specialist is possible at RP1 and RP2. (3) It is presumed that design work can and often does overlap across phases and that movement between phases is bi-directional and potentially iterative (e.g., P1 through P4). (4) “PT” refers to pretesting and “QA” refers to quality assessment.

- *Mode.* The selection of a data-collection mode (or modes, as the case may be) clearly has an impact 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.

The framework described above, though still very much a work-in-progress, has helped me in efforts to understand the role of evaluation methods within the context of the questionnaire design-and-evaluation process, and how each step of that process contributes to measurement error (and accuracy) during the survey administration phase (P5).<sup>5</sup> I find it difficult now to discuss evaluation methods out-of-context, that is, without making connections to early phases of the broader design-and-evaluation process (especially P1, P2 and P3). It is important to make these connections explicit and I try to do so when discussing methodological details and findings in section 5. For those readers whose primary interest is methodology, let me apologize in advance should you find my subsequent allusions to the framework distracting.

Before moving on to methodology, let us turn now to a brief discussion of the target questionnaire.

### 3. TARGET QUESTIONNAIRE: THE DISPLACED-WORKER 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. Manufacturing plants were closed, shifts were eliminated, and workers lost good-paying jobs. In an effort to assess the effects of these developments on the labor force, a small group of content specialists (labor economists) at the Bureau of Labor Statistics (BLS) set about to develop a questionnaire that would estimate the number of workers who were displaced from jobs. 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, the data it generated had utility for both internal and external users and, as a result, it has been administered biennially ever since (1984 through 2002).<sup>6</sup> 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). *Displaced workers* have been defined as follows:

“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)

---

<sup>5</sup> One modification to the framework presently under consideration is an expansion in the number of phases from six to eight; this would entail adding an evaluation phase after both the observation phase and the conceptualization phase. This modification would serve to underscore the crucial nature of these two early phases.

<sup>6</sup> In fact, the questions that comprise the DWS have almost always been followed by one or two other question modules that request information on such topics as job tenure and occupational mobility. To simplify the discussion, we chose to focus exclusively on the DWS questions in this paper.

### 3.2 Supplement Metadata: Interviewer Instructions, Concept Specification and the Displaced Worker Classification Algorithm

Several of the more important analytical items from the DWS appear in the Appendix, section A, along with associated metadata (e.g., concept specifications; data-entry instructions). This information was extracted from an instructional memorandum provided to all CPS interviewers prior to the February 2000 administration of the supplement (U.S. Bureau of the Census, 2000). In addition to this information, interviewers were instructed to refer to the CPS interviewing manual for definitions of key labor force concepts (e.g., work, job, business, layoff, and self-employment). The CPS “job” concept is defined in section B. The algorithm for classifying persons as displaced workers, a very important piece of metadata, can be found in section C. The reader is encouraged to review the contents of the Appendix. The development, dissemination and comprehension of such metadata are crucial for understanding the origins of measurement error and the interrelationships between its various sources.

### 3.3 Expert Review

In June 1995, the present author, serving as a questionnaire design specialist, was asked to review the displaced worker supplement (DWS) to identify potential sources of measurement error. The DWS comprises twenty-seven substantive items (earnings questions being excluded from the count); six of these items are involved directly in classifying persons as displaced workers (see Appendix, section C). The review also included twenty items from two other modules that were appended to the DWS (see footnote 6). The review, which was *not* based on a formal coding scheme (e.g., Lessler and Forsyth, 1996) but which was sufficient for the purposes intended, identified a number of potential problems with the supplement: (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 sponsor to authorize that quality assessment research be conducted in February 1996.

### 3.4 Key Supplement Questions: SD1 and SD2

Most of the evaluation data to be reviewed in this paper focuses on two key supplement items: SD1 and SD2 (see Table 2). 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; as a result, an evaluation of data (and metadata) associated with these items is crucial for detecting evidence of measurement error. A response of “no”, “don’t know” or “refused” to the first displacement question, SD1, moves the respondent out of the displacement series, and a response of “yes” takes the respondent to the second displacement question, SD2. Any answer categorized into one of the first three response options to SD2 takes the respondent through the remainder of the displacement series; however, not all persons who proceed down this path are classified as displaced. For example, persons who lost their jobs in the most recent year of the three-year reference period and who expect to be recalled within six months are not counted as displaced. An answer to SD2 that is categorized as “seasonal job completed”, “self-operated business failed”, or that an interviewer codes as “some other reason”, skips the respondent out of the displacement series.

**Table 2. Supplement Items SD1 and SD2 (Adults, Unweighted Data, 1996—2000)**

| 1996<br>[N=76,112] | 1998<br>[N=79,503] | 2000<br>[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?  |
|                    |                    |                    | 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<br>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.]   |

## 4. METHODOLOGY

The research conducted on the displaced-worker supplement during the period 1995-2000 (Esposito and Fisher, 1998; Esposito, 2000, 2001) is based on a multiple-method approach to questionnaire evaluation that was used in the early 1990s by researchers at the BLS and the Census Bureau to redesign the CPS (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). The collection of evaluative data from multiple sources using a variety of methods allows wording and sequencing problems to be identified, as well as problems in survey administration. 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; Cannell, Oksenberg 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. [As a point of information, two other methods were utilized at different points in time to evaluate the effectiveness of the DWS and to gather metadata on the supplement: An informal

expert review (see subsection 3.3) and several expert panels with content specialists (i.e., labor force economists with expertise using displaced-worker data).] General descriptions of the three principal evaluation methods are provided below. Phase-specific details pertaining to these methods will be provided later.

#### **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; Esposito and Hess, 1992), 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 Census Bureau's three centralized telephone centers. Senior staff supervisors from the telephone centers were asked to select a representative group of interviewers to serve as focus group participants—the principal criterion for selection being experience as an interviewer. Several days prior to administering the DWS, interviewers were given *log forms* on which to record any problems they may 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., supplement items 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 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 from 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, but these extended exchanges were not included in data summaries.

#### **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). With a survey methodologist (myself) assuming primary responsibility, 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).

**Table 3: Overview of Methods and Findings for Three Evaluation Phases (1996-2000)**

|                                | <b>Comments (C), Methodological Details (D) and Illustrative Findings (F)</b>  |
|--------------------------------|--|
| <b>Phase 1 (1996)</b>          | <ul style="list-style-type: none"> <li>▪ <b>C:</b> This phase can best be described as exploratory quality-assessment research. This initial evaluation focused on two supplement items, SD1 and SD2.</li> </ul>   |
| <i>Interviewer Debriefing</i>  | <ul style="list-style-type: none"> <li>▪ <b>D:</b> One focus group involving 10 telephone center interviewers.</li> <li>▪ <b>F:</b> 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).</li> </ul>  |
| <i>Interaction Coding</i>      | <ul style="list-style-type: none"> <li>▪ <b>D:</b> 52 person interviews coded (behavior coding).</li> <li>▪ <b>F:</b> 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).</li> </ul>  |
| <i>Respondent Debriefing</i>   | <ul style="list-style-type: none"> <li>▪ <b>D:</b> Debriefing questionnaire consisting of 8 response-dependent probe questions.</li> <li>▪ <b>F:</b> 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).</li> </ul>   |
| <b>Phase 2 (1998)</b>          | <ul style="list-style-type: none"> <li>▪ <b>C:</b> Relative to the quality assessment work conducted in 1996, this second phase was far more comprehensive. Again, the evaluation focused on SD1 and SD2.</li> </ul>   |
| <i>Debriefing Interviewers</i> | <ul style="list-style-type: none"> <li>▪ <b>D:</b> 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.</li> <li>▪ <b>F:</b> 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.</li> </ul> |

- |                               |  |
|-------------------------------|--|
| <i>Interaction Coding</i>     | <ul style="list-style-type: none"> <li>▪ <b>D:</b> 145 person interviews coded (behavior coding).</li> <li>▪ <b>F:</b> 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).</li> </ul>  |
| <i>Respondent Debriefing</i>  | <ul style="list-style-type: none"> <li>▪ <b>D:</b> Debriefing questionnaire consisting of 22 response-dependent probe questions.</li> <li>▪ <b>F:</b> 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).</li> </ul>                    |
| <b>Phase 3 (2000)</b>         | <ul style="list-style-type: none"> <li>▪ <b>C:</b> 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).</li> </ul>  |
| <i>Interviewer Debriefing</i> | <ul style="list-style-type: none"> <li>▪ <b>D:</b> Two focus groups involving 22 telephone center interviewers.</li> <li>▪ <b>F:</b> 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).</li> </ul> |
| <i>Interaction Coding</i>     | <ul style="list-style-type: none"> <li>▪ <b>D:</b> 131 person interviews coded (behavior coding).</li> <li>▪ <b>F:</b> 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).</li> </ul>   |
| <i>Respondent Debriefing</i>  | <ul style="list-style-type: none"> <li>▪ <b>D:</b> Debriefing questionnaire consisting of 11 response-dependent probe questions.</li> <li>▪ <b>F:</b> Evidence of a possible displaced-worker undercount of 29 percent (false negatives). 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).</li> </ul>  |
- 

## 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).<sup>7</sup>

<sup>7</sup> I wish to acknowledge Sylvia Fisher’s valuable contributions in implementing this initial research effort.

### 5.1.1 Methodological Details and Findings

Relative to subsequent phases, the scope of this initial evaluation was limited. With respect to the interviewer debriefing, one focus group was conducted with ten CPS interviewers serving as research participants. Examples of questions used to debrief interviewers on SD1 and SD2 appear in Table 4. With respect to SD1, interviewers noted the following: (1) Some respondents seemed to have difficulty with the distinction intended (but not made explicit) between losing a job and leaving one; (2) some respondents interpreted the phrase “or another similar reason” too broadly; and (3) some persons reported to have lost a job actually may not have had a job to lose (e.g., on-call substitute teachers).<sup>8</sup> With respect to SD2: (1) most interviewers never read the “read if necessary” statement—there would have been no reason to do so unless the respondent volunteered relevant information or hesitated or requested clarification when answering SD1; and (2) many interviewers were not aware that the “other” precode skips the case out of the displacement series—as a result, some felt it was their responsibility to correct answers to SD1 when they encountered cases that obviously were not displacements.

**Table 4. Examples of Interviewer Debriefing Questions (Phase One, 1996)**

---

|            |  |
|------------|--|
| <b>SD1</b> | <ul style="list-style-type: none"> <li>▪ Did you have difficulty reading this question in its entirety before respondents provided an answer?</li> <li>▪ Did the respondents have difficulty with the concept of “lose a job or leave one”?</li> <li>▪ Were respondents able to distinguish the four response options presented to them? If not, what confusions or misconceptions did they report?</li> <li>▪ Was the phrase “or another similar reason” causing any problems for respondents?</li> <li>▪ How clear were instructions on classifying a response so it could be matched to one of these four options?</li> </ul>   |
| <b>SD2</b> | <ul style="list-style-type: none"> <li>▪ Did you have difficulty reading this question in its entirety (i.e., all 6 response options)?</li> <li>▪ 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”?</li> <li>▪ How frequently did you read the READ AS NECESSARY statement?</li> <li>▪ 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?</li> <li>▪ Were there any additional reasons offered by respondents for their job loss not available in the current list? If yes, what were they?</li> </ul> |

---

With regard to behavior coding, twenty-three CPS household interviews were monitored and interviewer-respondent exchanges for fifty-two person interviews were coded (see Table 5); each household interview typically involved one or more *person interviews*—data were collected for each CPS-eligible member of the household. Coded data suggest that interviewers experienced some difficulty reading SD1 as worded and 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.

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

<sup>8</sup> A telephone-center staff supervisor, not an interviewer, provided the last comment in the series.

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). Recall that 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 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 to 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=1032, numerator), and then divided by the appropriate denominator (i.e., base equals 4211, the sum of precodes 1, 2 and 3 for SD2), these debriefing data suggest a displaced-worker *undercount* of approximately 25 percent—that is, 25 percent *over and above* the official estimate. 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).

**Table 5. Behavior Coding Data for Selected Items (1996-2000)**

| Phase               | Item(s)                | Interviewer Codes |                 | Respondent Codes |                 |               |                 |
|---------------------|------------------------|-------------------|-----------------|------------------|-----------------|---------------|-----------------|
|                     |                        | E                 | MC              | AA               | IA              | RC            | INT             |
| <b>One (1996)</b>   | SD1                    | 65%<br>(33/51)    | 16%<br>(8/51)   | 88%<br>(42/48)   | 2%<br>(1/48)    | 8%<br>(4/48)  | 19%<br>(9/48)   |
|                     | SD2                    | 29%<br>(2/7)      | 57%<br>(4/7)    | 67%<br>(4/6)     | 33%<br>(2/6)    | 0%<br>-       | 17%<br>(1/6)    |
| <b>Two (1998)</b>   | SD1                    | 71%<br>(96/135)   | 13%<br>(18/135) | 88%<br>(119/135) | 10%<br>(13/135) | 1%<br>(1/135) | 25%<br>(34/135) |
|                     | SD2                    | 0%<br>-           | 72%<br>(13/18)  | 56%<br>(10/18)   | 28%<br>(5/18)   | 0%<br>-       | 39%<br>(7/18)   |
| <b>Three (2000)</b> | SD1                    | 69%<br>(82/119)   | 18%<br>(22/119) | 93%<br>(110/118) | 5%<br>(6/118)   | 0%<br>-       | 13%<br>(15/118) |
|                     | SD2                    | 29%<br>(4/14)     | 43%<br>(6/14)   | 60%<br>(6/10)    | 40%<br>(4/10)   | 0%<br>-       | 0%<br>-         |
|                     | SDB3                   | 93%<br>(110/118)  | 3%<br>(4/118)   | 98%<br>(115/117) | 0%<br>-         | 2%<br>(2/117) | 5%<br>(6/117)   |
|                     | [SDB2A/B +<br>SDB5A/B] | 74%<br>(14/19)    | 21%<br>(4/19)   | 74%<br>(14/19)   | 26%<br>(5/19)   | 0%<br>-       | 16%<br>(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 greater 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 time 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 | <p>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?</p> <p><i>Rationale:</i> 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).</p>  |
| SDB2 | <p>Earlier in this interview, when answering questions about the job you had lost or left between January 1993 through December 1995, were you answering the questions based on the job that you had held for the longest time?</p> <p><i>Rationale:</i> 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 <i>from which a displacement occurred</i>), reporting errors were possible. SDB2 was an attempt to quantify that error.</p>  |
| SDB3 | <p>Did you lose that job or did you leave that job?</p> <p>[<i>Rationale:</i> 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.]</p>   |
| SDB4 | <p>During the time period spanning January 1993 through December 1995, did you leave a job or retire from a job?</p> <p><i>Rationale:</i> 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).</p>   |
| SDB5 | <p>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., <i>company or plant had insufficient work; was downsizing or restructuring; was filing for bankruptcy</i>).</p> <p><i>Rationale:</i> 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.</p> |

---

### 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 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. In fact, 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 little documentation available on the original 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 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*. Though somewhat vague, the information provided to interviewers appears to confirm this impression; however, all such cases are skipped out of the displacement series (see SD2, precode 6).<sup>9</sup> Third, various aspects of the displacement concept, as understood by content specialists, 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* versus 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 twenty 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 phases of the design-and-evaluation process described earlier (see Table 1, phases P1 through P3) and provide a sense of how such problems contribute to measurement error during the survey administration phase (P5). The intent of pretesting (P4) 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 (P6) conducted in 1996 provided direct and indirect evidence of measurement error (at P5), we always need to exercise care in interpreting evaluation findings. For example, after reading a draft of the evaluation report, one astute reviewer, a content specialist, noted that very little effort had been expended in identifying false positives—a valid observation.<sup>10</sup> An attempt to correct this imbalance was made in subsequent research. We should also keep in mind that 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. 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

---

<sup>9</sup> Regarding “similar reasons”, interviewers were provided with the following information: “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 counterintuitive, the decision to skip “some other reason” entries (SD2, precode 6) out of the displacement series actually *reduced* measurement error. Including those cases would have generated about four times as many false positives (81%) relative to false negatives (19%). Most of the precode-6 entries do not constitute displacements.

<sup>10</sup> We thank Anne Polivka for offering this comment and for providing other incisive feedback. We might note here that, while false positives have the beneficial effect of offsetting false negatives (thus moderating the magnitude of a presumed undercount), they still represent measurement error and may have significant negative consequences for other estimates (e.g., miscalculations of outcome measures, such as earnings disparities between an old and a new job). In evaluation research, every effort should be made to identify both false negatives and false positives.

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 thirty-four 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 (for a review, see Table 4). With respect to SD1, most of the substantive problems identified during phase one were again observed here. Other substantive problems were also noted such as: uncertainty regarding transitions involving self-employed persons and persons who worked for temporary staffing agencies; ambiguity regarding the meaning of various technical terms (e.g., “insufficient work”; “fired”); and uncertainty regarding how to deal with unusual work arrangements and involuntary downgrades in job status. Interviewers also mentioned a variety of pragmatic problems (e.g., relevance of questions for the long-term retired and disabled; response difficulty for parents who were serving as proxies for their older, job-hopping children). With respect to SD2, interviewers mentioned similar problems: ambiguity regarding the meaning technical terms (e.g., “seasonal job”; “layoff”); reports of alleged employer deception; and misreporting job loss. Interviewers also mentioned a variety of pragmatic problems with SD2: rambling stories about separation events that interviewers are left to decipher and code; default use of the “some-other-reason” precode whenever there was doubt as to the accuracy of the response-to-precode matching process; awkwardness of the embedded transition statement; and interruptions. Many of the problems noted above 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 also asked interviewers to rate these items in terms of how difficult they thought it was for respondents to provide adequate answers (see Table 7); this rating task was only performed when interviewers spontaneously identified the question as problematic during the problem-identification stage of the focus group. 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 between groups.

**Table 7. Interviewer Ratings for Supplement Items SD1 and SD2 (Phase Two, 1998)**

| TC Location    |      | 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)*

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 like the first phase, interviewers struggled with the wording of both SD1 and SD2, especially the latter, and respondents experience 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 twenty-two 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 to SD1 but for whom responses to subsequent debriefing questions (SDB17 and SDB20) suggested that they may 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 3670, the sum of precodes 1, 2 and 3 for SD2), these debriefing data suggest a displaced-worker *undercount* of approximately 20 percent—that is, 20 percent *over and above*

the official estimate.<sup>11</sup> As the data suggest, about three-tenths of this error is traceable to path one (SD2, precode 6), and the remainder is attributable to path two (SD1=no).

**Table 8. Examples of Respondent Debriefing Questions (Phase Two, 1998)**

---

|       |   |
|-------|---|
| SDB1  | <p>Earlier you told me that me that you had lost or left a job in the past three calendar years [<i>fill with displacement reason from SD2</i>]. Did you lose that job or did you leave that job?</p> <p><i>Rationale:</i> 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 <i>job leavers</i> to specific follow-up probes.</p>  |
| SDB2  | <p>Was the job you (fill: “left”) a temporary job, that is, a job that was supposed to last only for a limited time or until the completion of a project?</p> <p><i>Rationale:</i> 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.</p>  |
| SDB3  | <p>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 twenty-two response precodes, seven <i>employer-related reasons</i> (e.g., business closed down; restructuring; insufficient work; position/shift abolished) and fifteen <i>personal reasons</i> (e.g., did not like job or boss; better job; not enough pay; own illness/injury; fired; school/training).</p> <p><i>Rationale:</i> Generally speaking, to determine if the person lost or left a job involuntarily (i.e., one of the employer-related reasons) or involuntarily (i.e., one of the personal reasons). This item was useful for identifying potential <i>false negatives</i>.</p> |
| SDB3Z | <p>Did you ever return to work for that employer, for even a short period of time?</p> <p><i>Rationale:</i> For persons reported to have <i>lost, left, or retired from</i> 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 <i>false positives</i> (e.g., persons who returned to work for their former employers, presumably doing the same work and not subsequently displaced again).</p>  |
| SDB17 | <p>During the period January 1995 through December 1997, did you leave a job or lose a job for any reason?</p> <p><i>Rationale:</i> 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).]</p>   |
| SDB20 | <p>What is the MAIN reason you are no longer working at that job? [Note: This item had twenty-two response precodes, seven <i>employer-related reasons</i>) and fifteen <i>personal reasons</i> (see SDB3 for examples).</p> <p><i>Rationale:</i> 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). This item was useful for identifying potential <i>false negatives</i>.</p>  |

---

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

<sup>11</sup> The undercount for phase one was estimated to be 25 percent. The phase-two figure (20 percent) may be lower, in part, because persons who had lost/left temporary jobs (“SD1=no” path) were not asked to provide a reason for their separation (i.e., not asked SDB20)—in retrospect, a poor decision, because some of these workers probably were displaced from a job. In phase one, we did not identify temporary jobs as part of the debriefing process.

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 and 11.5 percent, respectively. Some content 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 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. For example, interviewers again reported problems with key concepts (e.g., job, insufficient work, temporary versus permanent separations) and with administrative tasks (e.g., collecting data from elderly respondents; proxy reporting for older children). 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).

The implications of these findings were not lost on program managers. Meetings with internal and external subject-matter experts served to clarify various conceptual issues and determine user needs. The supplement sponsor authorized steps 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 regular basis to review evaluation data and to discuss questionnaire design issues (e.g., concepts, supplement 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). Using the framework provided earlier (see Table 1), these redesign activities might best be classified as work within the conceptualization, operationalization and evaluation phases of the questionnaire-*redesign*-and-evaluation process (RP2, RP3 and RP4 phases, respectively).<sup>12</sup>

<sup>12</sup> The careful reader may have noticed that we have not alluded to RP1 activities in this 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,

### 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 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., 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 the administration of the 2002 DWS, two forms of evaluation were taking place simultaneously: limited *quality-assessment work* on the DWS and *pretesting work* (field based) on several items under consideration for the new JSS.

#### 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 twenty-two 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 (see Table 4). The questions used to debrief interviewers on the prospective items for the new job-separation supplement can be found in Table 9 (see bullets under items SDB1 through SDB6). Though we had expected little new from interviewers with respect to SD1 and SD2, such was not the case. For example, interviewers reported that some respondents were uncertain about how to answer these items: when an employee's company merged with another or changed hands; or when a person was transferred, usually involuntarily, to a different place of work. Some interviewers were not sure how to code answers: when workers left jobs that had a predetermined endpoint (e.g., construction work); when an employee left a job as a result of a cut in hours; or when an employee might have been prodded into quitting a job. Most of these examples refer to fundamental conceptual issues: What constitutes a job? When is "leaving a job" indistinguishable in principle from losing a job? Unfortunately, these issues do not vanish when we turn our attention to items that are being considered for the new supplement.

In the JSS, item SDB3 is being considered as the supplement's 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 with an employer 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 (e.g., SDB5A and SDB5B). Some of the problems noted above for SD1 and SD2 are relevant for SDB3 as well. However, other examples of problematic situations surfaced here. For example, some respondents were uncertain how to answer this question when an employee left one position within the company to take another (e.g., "switched accounts"). Some interviewers were not sure how to code answers: when workers reported moving from job to job as part of their trade (e.g., plumbers, electricians); or when the person's job involved work for different clients or employers (e.g.,

---

content specialists at the BLS did have a substantial amount of information/data available that might qualify as *indirect* observations. That information/data came from four sources: (1) evaluation data from the present research effort (phases one and two)—particularly respondent debriefing data (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 (e.g., Mass Layoff Statistics; Job Opening and Labor Turnover Survey).

**Table 9. Examples of Interviewer Debriefing Questions (Bullets) for Selected Respondent Debriefing Items (Phase Three, 2000)**

---

|              |  |
|--------------|--|
| <b>SDB1</b>  | <p>Earlier you told me that you had lost or left a job during the period 1997 through 1999 because (<i>fill with displacement reason from SD2</i>). 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.]</p> <ul style="list-style-type: none"> <li>▪ Did any respondents appear to have difficulty understanding what was meant by the term “job”?</li> <li>▪ Did any respondents appear to have difficulty understanding the difference between losing a job or leaving a job?</li> <li>▪ Did you notice any special problems that proxy respondents might have had in answering this question?</li> </ul>  |
| <b>SDB2A</b> | <p>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: Twenty substantive response options were provided, seven referred to <i>employer actions</i> (e.g., employer was: closing down company; moving, merging or selling company; cutting back jobs; downsizing, reorganizing, or outsourcing jobs) and thirteen referred to <i>personal reasons/actions</i> (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 .)]</p> <ul style="list-style-type: none"> <li>▪ Did you have difficulty reading this question in its entirety?</li> <li>▪ Did any respondents have difficulty identifying the MAIN reason why the target person is no longer working for her/his former employer?</li> <li>▪ Did you experience any difficulty matching respondent’s answers to the available response options? If YES: What types of coding problems did you encounter?</li> <li>▪ 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)?</li> </ul> |
| <b>SDB2B</b> | <p>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.]</p>   |
| <b>SDB3</b>  | <p>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.]</p> <ul style="list-style-type: none"> <li>▪ [Note: See first two bullets/questions listed for SDB1.]</li> <li>▪ Did any respondents appear to have difficulty understanding the difference between a full-time job and a part-time job?</li> </ul>  |
| <b>SDB3B</b> | <p>How many jobs, total, did you lose or leave during 1999?</p> <ul style="list-style-type: none"> <li>▪ [Note: See first bullet/question listed for SDB1.]</li> <li>▪ Did any respondents have difficulty recalling how many jobs the person had lost during 1999?</li> </ul>   |
| <b>SDB4</b>  | <p>[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?</p> <ul style="list-style-type: none"> <li>▪ [Note: See first two bullets/questions listed for SDB1.]</li> </ul>   |
| <b>SDB5A</b> | <p>[Note: Same wording as SDB2A, but asked of respondents who answered “no” to SD1.]</p>   |
| <b>SDB5B</b> | <p>[Note: Same wording as SDB2B, but asked of respondents who answered “no” to SD1.]</p>   |
| <b>SDB6</b>  | <p>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?</p> <ul style="list-style-type: none"> <li>▪ Did any respondents appear to have difficulty understanding the concept of a “temporary job”?</li> </ul>   |

---

freelance workers). In the proposed classification algorithm for the JSS, to be considered displaced, persons who *leave* jobs may have to satisfy additional conditions (e.g., written advance notice) relative to persons who *lose* jobs, so this distinction will have to be established within the supplement (see SDB4 for wording). Some of the problems with this distinction have been discussed above. Another problem mentioned by interviewers had to do with temporary staffing jobs. Some proxy respondents do not seem to realize that “temp workers” are employees of temporary staffing agencies, and not the companies to which they are assigned. 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 items SDB2A and SDB2B for wording). Operationally, this set of questions can pose challenges for interviewers who first have to extract the essence of a respondent’s sometimes lengthy and/or vague answer, and then find a match for that answer among twenty 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 like the first two phases, interviewers struggled with the wording of both SD1 and SD2, especially the latter, and respondents experience 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 job-separations supplement (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 eleven 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.<sup>13</sup> 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

---

<sup>13</sup> Unlike prior estimates of false negatives based on *three-year* reference periods (e.g., Esposito and Fisher, 1998), the specific wording of debriefing item SDB3 made it possible to produce an estimate for the *last year* (1999) of the present three-year reference period (1997-1999) (Esposito, 2001, pp. 18-24). Interestingly, data from debriefing item SDB6 suggest that approximately 11 percent of these potential false negatives—that is, 11 percent of the estimated 29 percent—worked at jobs that were characterized as temporary; to the extent that their jobs truly were temporary, these individuals probably should not be counted among the displaced. In other words, a more accurate estimate of false negatives may actually be 26 percent [i.e., 29 percent minus 3 percent (11 x 29, see above)].

regarding this puzzling finding (e.g., overlooking a marginal or a part-time job; entry errors; a fatigue effect; a respondent conditioning effect; purposeful misreporting; evaluation methodology 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 test 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 (hypothesis one) or because they more than one job (hypothesis two), data from debriefing items SDB7 and SDB8 do not provide strong support for either hypothesis.

**Table 10. Behavior-Coding Protocols Containing Possible Classification Errors (Phase Three, 2000)**

| ITEM   | RESPONSE  | INTERVIEWER ENTRY  |
|--|---|--|
| <b>Case 49 Protocol (Second Person in Household)</b> |   |  |
| SD1  | The respondent answered, yes, that he had <i>left</i> a job.                  | Precode 1 (yes).   |
| SD2  | The respondent answered that he was “terminated” from a job with Clark Oil.   | 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 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 respond by saying they left a job. But such self-presentation strategies can have an impact on displacement estimates if persons who *leave* jobs have to satisfy certain conditions (e.g., written advance notice) that persons who lose jobs do not have to satisfy to be classified as displaced.

**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).

**Case 56 Protocol (First Person in Household)**

|       |   |   |
|-------|---|---|
| SD1   | The respondent said her division/department was sold.   | Precode 1 (yes).  |
| SD2   | No response possible. Interviewer never read any part of SD2; she simply entered what she thought was the appropriate code. | Precode 6 (other).  |
| Note: | [Intervening questions omitted.]  |   |
| SDB1  | Lost job.   | Precode 1 (lost job).   |
| SDB2A | “You can call it a merger.”   | [Unknown, but precode 2 would have been appropriate (employer was moving, merging, or selling the company, plant or office).] |

**Comments (Case 56):** There is no ambiguity about this case; the person should have been classified as displaced and, based on the way his response were coded, that will not happen (i.e., false negative).

**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. 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 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 box; these respondent debriefing items did not have to shoulder the full burden of identifying displaced workers independently. Second, data from behavior coding and/or 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. Among the principal objectives of a new supplement on job separations is that it produce an accurate annual count of displaced workers and an accurate annual displacement rate, given relatively fixed cost and operational constraints.

The research described above suggests that there are two potentially useful approaches to gathering data on displacement and other job separations—a *free response approach* (asking an field-coded, open question with displacement reasons provided as response precodes) and a *direct approach* (asking a closed question with displacement reasons explicitly mentioned in the question stem). While prior research has provided useful metadata on both of these approaches, it does not provide an experimental basis for choosing between the two. To generate such data, we hope to conduct a split-panel test when research funds become available; such a field test would also provide the opportunity to evaluate other items on the current draft of the job-separations supplement (see Table 1, phase RP4).

## 6. GENERAL DISCUSSION

In the introduction to this paper, four ways in which this research might contribute to questionnaire-evaluation practice and theory were noted. I would like to revisit those aspirations in this closing section.

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

While no single evaluation effort can be expected to identify all problems associated with a given questionnaire, the papers presented at this conference provide sufficient evidence to conclude that questionnaire evaluation research is indeed effective at identifying real problems. 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 supplement item 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, other than poor question design, we are unable to offer a more compelling explanation for these false negatives. 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, researchers learn which types of questions do not work; we also learn how to design 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”. As an example of the second point, the verbatim (other/specify) 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 job separations supplement.

Many readers will be concerned, understandably, about the costs associated with iterative questionnaire evaluation research. On that concern, consider three points. First, costs can usually be controlled by limiting the scope of research and, to the extent possible, by assigning work to in-house staff. Much of the work reported above could have been accomplished by one survey methodologist—in collaboration, of course, 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 (e.g., technology use). And third, as an alternative to periodic evaluations, one could ask: What are the costs associated with collecting and disseminating poor-quality data? Though many survey organizations are presumably working hard to improve data quality, at present there simply does not appear to be an efficient way to collect high-quality survey data inexpensively.

## **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 the 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.

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. While increasing the labor component may increase the yield of a particular technique, 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 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), I suspect the scores and values associated with these techniques would have been different as well.

We should acknowledge, as many practitioners have (see citations in footnote 3), that 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 aim. 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

**Table 11. A Subjective Assessment of Productivity, Information Yield and Labor Investment Associated with Four Questionnaire Evaluation Techniques (Phase Two, 1998)**

| <i>Method/Technique</i>               | <b>P</b> | <b>Y</b> | <b>L</b> | <b>Comments</b>   |
|---------------------------------------|----------|----------|----------|---|
| <b><i>Interviewer Debriefing</i></b>  |          |          |          |   |
| ▪ Focus Group                         | 1.25     | 5        | 4        | <ul style="list-style-type: none"> <li>▪ Qualitative data: Retrospective and subject to situational effects (e.g., group dynamics).</li> <li>▪ Useful for identifying conceptual and operational problems.</li> <li>▪ Sample of interviewers not representative of population.</li> <li>▪ Provides no quantitative basis for estimating measurement error.</li> </ul>   |
| ▪ Rating Form                         | 1.00     | 1        | 1        | <ul style="list-style-type: none"> <li>▪ Descriptive quantitative ratings data: Retrospective and potentially contaminated if interviewers talk about items.</li> <li>▪ Useful in identifying differences among interviewers, but sample of interviewers not representative of population.</li> <li>▪ Minimal labor on part of researcher.</li> <li>▪ Provides no quantitative basis for estimating measurement error.</li> </ul>   |
| <b><i>Interaction Coding</i></b>      |          |          |          |   |
| ▪ Behavior Coding (live)              | 1.00     | 3        | 3        | <ul style="list-style-type: none"> <li>▪ Descriptive quantitative data and some qualitative data.</li> <li>▪ Useful in detecting possible problems with specific items, but not necessarily useful in identifying solutions.</li> <li>▪ Useful for comparative analyses (open vs. closed questions)</li> <li>▪ Relatively objective/unbiased.</li> <li>▪ Sample of interviewers and respondents not fully representative of their respective populations.</li> <li>▪ Live coding more susceptible to error and omissions than other coding strategies (e.g., coding from audiotapes)</li> <li>▪ Provides no quantitative basis for estimating measurement error.</li> </ul>   |
| <b><i>Respondent Debriefing</i></b>   |          |          |          |   |
| ▪ Response-dependent Follow-up Probes | 1.50     | 9        | 6        | <ul style="list-style-type: none"> <li>▪ 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.</li> <li>▪ Qualitative data: "Other-specify" precodes provide quasi-ethnographic data.</li> <li>▪ Respondent sample fairly representative of population.</li> <li>▪ Adds to respondent burden in some cases.</li> <li>▪ Labor intensive for content and design specialists.</li> <li>▪ 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.</li> </ul> |

**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.

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 (myself) than with the technique—was that only interviews with English-speaking respondents could be monitored and coded.

As the discussion above suggests, all survey evaluation techniques have weaknesses associated with them. 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 (Esposito and Rothgeb, 1997). The multiple-method evaluation strategy is not about discovering “truth”, which I would assume is unattainable in all but simulations or trivial cases. It is more about triangulating on an *understanding* of what might be problematic regarding a particular questionnaire. 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**

This case study would appear to provide motive for offering the following (over-)generalization: *Inadequacies in conceptual specification lead inescapably to measurement error* (and, with the use of appropriate methods, that error can be quantified in approximate terms).<sup>14</sup> Though certainly not a novel observation (Federal Committee on Statistical Methodology, 1988; Groves, 1989; Hox, 1997; Martin, 1987; Turner and Martin, 1984, Chapter 7), it is one to which we must pay more than lip service when designing questionnaires. In this paper, considerable space has been devoted to providing metadata (e.g., key definitions; interviewer documentation; the displaced-worker classification algorithm) and examples that relate specifically to these conceptual issues. It is important to see the connections among these metadata and actual questionnaire content, and to relate these connections (or lack thereof) to *manifestations* of measurement error—be they direct and indirect. One of the easiest examples to follow relates to the measurement error associated with the “another similar reason” wording in supplement items SD1 and SD2. If you have not done so already, please review the interviewer instructions for these two items (see Appendix, section A). Carefully read the descriptive information provided

---

<sup>14</sup> 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.” Several sources of specification error are noted: “(1) inadequately specified uses and needs, (2) inadequately specified concepts, and (3) inadequately specified data elements.”

for “similar reasons” (SD1) and for “some other reason” (SD2). Does this information seem clear? Using this descriptive information and that provided for the other response precodes, the CPS definition of “job” (section B) and the definition of displaced workers provided earlier (see subsection 3.1), how would you have coded the following verbatim responses: “company merged with another company”, “another bank bought out bank that person was working for”, “laid off permanently”, “employer cut person’s hours”, “employer sold business—new owner didn’t need workers”, “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”, “employer was trying to get rid of experienced staff”?<sup>15</sup>

Take a look at the DWS classification algorithm (section C)—all of the key questions are described in section A. Doing so, you may note that persons out of work for as long as twelve months (SD3), who have *an expectation* that they will return to work within another six months (SD4) are not counted among the displaced; that is, none of these persons would be categorized as displaced, even though an undetermined percentage may never return to work for their former employers. Does that seem like a reasonable specification? Recall that the American economy of the early 1980s was very different relative to the economy we have now. The former was dominated by manufacturing (e.g., automobiles and steel), the latter, by services. Take another look at the wording of SD1 and the precodes of SD2: Do they reflect the current economy? Review the verbatim entries provided above (SDB3 and SDB20): Do these examples reflect the economy of the early 1980s?

The objective of this discussion is *not* to disparage the displaced worker supplement, its sponsor or its designers; this supplement actually served its original purpose fairly well. The objective is to drive home two important 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**

It is my hope that the framework described earlier (see section 2) will be useful in addressing theoretical issues and in solving practical problems. 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

---

<sup>15</sup> These examples were drawn from the other-specify precodes of debriefing items SDB3 and SDB20 (phase 2).

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 (P4) or quality assessment (P6), or of dismissing such evaluations altogether? What would be the implications of not involving design specialists in the operationalization phase or of not involving content specialists in the evaluation process?

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 P4). Consider column one: Pretesting work taking place at C<sub>41</sub> could cause content specialists to revisit C<sub>11</sub> and C<sub>21</sub> activities (observation and conceptualization, respectively), because recursive movement is possible between P1 and P4. In principle, it is not possible for activities at C<sub>61</sub> (quality-assessment work) to affect activities at C<sub>11</sub>, because the P5 phase (survey administration) acts as a barrier to recursivity. However, activities at C<sub>61</sub> may have an effect on subsequent observational activities of a content specialist (C<sub>R11</sub>) even though no formal redesign effort may yet be underway (see below). We might also consider the interrelationships among various sources of error that exist during any one phase of the questionnaire design-and-evaluation process (e.g., P5). A review of the behavior coding protocols in Table 10 provides some insights into the nature of these interrelationships. For example, had the survey sponsor updated supplement instructions and provided 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 supplement designers done a better job with the wording of SD1 and SD2, 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* change in a target domain (e.g., change that threatens the integrity of a 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 these terms may increase measurement error to some extent. Do such changes in terminology represent semantic/superficial change or significant 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., questionnaire adjustments; redesign activities).

A few thoughts in closing: I believe that minimizing measurement error is an *ongoing process*. There would appear to be no single method or technique capable of identifying all questionnaire-related problems, no quick-fix methodology capable of patching the gaping conceptual holes characteristic of inadequate foundation work, and no individual or isolated group or single administration mode capable of guaranteeing accurate survey data. The quality assurance process starts with observation, and periodic observations/evaluations of surveys with recognized societal importance 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. Insofar as you are already a part of this process, I encourage you to adopt the organizational framework described herein as your own and utilize it in whatever way seems reasonable to improve our understanding of the questionnaire design-and-evaluation process.

## REFERENCES

- Abraham, K.A (1996). "Ensuring Quality in the Data Collection Process," Commissioner's Order No. 2-96, Washington, DC: 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.
- 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*, U.S. Bureau of the Census, 361-374.
- Cannell, C.F., Miller, P.V. and Oksenberg, L. (1981). "Research on Interviewing Techniques," in S. Leinhardt (ed.), *Sociological Methodology*, San Francisco: Jossey-Bass, 389-437.
- Cannell, C. and Oksenberg, L. (1988). "Observation of Behavior in Telephone Interviews," in R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicolls, II, and J. Waksberg (eds.), *Telephone Survey Methodology*, New York: Wiley, 475-495.
- Cannell, C., Oksenberg, L., Kalton, G., Bischooping, 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.

- Czaja, R. and Blair, J (1996). *Designing Surveys: A Guide to Decisions and Procedures*, Thousand Oaks, CA: Pine Forge Press.
- 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 T.J. DeMaio (ed.), *Approaches to Developing Questionnaires*, Statistical Policy Working Paper 10, Washington, DC: Office of Management and Budget, 119-136.
- Dippo, C. and Sundgren, B. (2000). "The Role of Metadata in Statistics," *Proceedings of the Second International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, 909-918.
- Dippo, C.S., Conrad, F.G. and Gillman, D.W. (2000). "Metadata and Data Quality", UN/ECE Work Session on Statistical Metadata, Working Paper Number 5, Washington, DC, USA, 28-30 November 2000.
- Esposito, J.L. (2001). "An Evaluation of Prospective Questions for a New CPS Supplement on Job Separations and Employee Tenure: Consolidated Report", Unpublished Report, Washington, DC: U.S. Bureau of Labor Statistics.
- Esposito, J.L. (2000). "Evaluating the Displaced-Worker/Job-Tenure Supplement to the CPS: An Illustration of Multimethod Quality-Assessment Research", *Statistical Policy Working Paper 30*, Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, 411-420.
- Esposito, J.L., and Fisher, S. (1998). "A Summary of Quality Assessment Research Conducted on the 1996 Displaced-Worker/Job-Tenure/Occupational-Mobility Supplement" *BLS Statistical Note Series*, Number 43, U.S. Bureau of Labor Statistics: Washington, DC.
- Esposito, J.L., and Rothgeb, J.M. (1997). "Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment", in L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, 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.
- Esposito, J.L., and Hess, J. (1992), "The Use of Interviewer Debriefings to Identify Problematic Questions on Alternate Questionnaires," paper presented at the Annual Meeting of the American Association for Public Opinion Researchers, St. Petersburg, FL.
- 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 P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, 393-418.
- Fowler, F.J., and Mangione, T.W. (1996). *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 N. Schwarz and S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, San Francisco: Jossey-Bass, 15-36
- Gerber, E. (1999). "The View from Anthropology: Ethnography and the Cognitive Interview," in M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds.), *Cognition and Survey Research*, New York: Wiley, 217-234.
- Groves, R.M. (1996). "How Do We Know What We Think They Think Is Really What They Think?," in N. Schwarz and S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, 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, Luxembourg, 6-8 March 2002.
- Hess, J.C., and Singer, E. (1995). "The Role of Respondent Debriefing Questions in Questionnaire Development," *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 1075-1080.
- Hox, J.J. (1997). "From Theoretical Concept to Survey Question," in L.Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, New York: Wiley, 47-69.
- 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.
- Lessler and Forsyth, 1996. "A Coding System for Appraising Questionnaires," in N. Schwarz and S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, San Francisco: Jossey-Bass, 259-291.
- Martin, E. (1987). "Some Conceptual Problems in the Current Population Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 420-424.
- Maynard, D.W., and Schaeffer, N.C. (2002). "Standardization and Its Discontents," in D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds.), *Standardization and Tacit Knowledge*, New York: Wiley, 3-45.
- 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 M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (eds.), *Cognition and Survey Research*, 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 P.V. Marsden (ed.), *Sociological Methodology*, Volume 24, 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 Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 46-55.
- Schuman, H. (1966). "The Random Probe: A Technique for Evaluating the Validity of Closed Questions," *American Sociological Review*, 218-222.
- Shepard, J. and Vincent, C. (1991). "Interviewer-Respondent Interactions in CATI interviews," *Proceedings of the Census Bureau's 1991 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 523-536.
- Sirken, M.G., Herrmann, D.J., Schechter, S., Schwarz, N., Tanur, J.M., and Tourangeau, R. (eds.) (1999). *Cognition and Survey Research*, New York: Wiley.
- 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.
- Schwarz, N., and Sudman, S. (eds.) (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, 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 T. Jabine, M.L. Straff, J.M. Tanur and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, 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.
- Tucker, C., Bloxham, J., Bowie, C., Esposito, J.L., Kojetin, B., Kostanich, D., Miller, S., Polivka, A., Robison, E., and Stump, M. (1997). "Improving Field Tests," Unpublished Report, Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics, and U.S. Department of Commerce, Bureau of the Census.
- Turner, C.F., and Martin, E. (1984). *Surveying Subjective Phenomena*, Volume 1, New York: Russell Sage Foundation.

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

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 Labor Statistics (1994). "The Bureau of Labor Statistics Quality Measurement Model," Washington, DC: U.S. Department of Labor.

QDET/Contributed/JLE-QDET 111902

**Suggested Citation:**

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, 14-17 November 2002, Charleston, SC.

## APPENDIX

### Section A: Specifications for Key Supplement Items (Bureau of the Census, 2000)

#### Item by Item Instructions for Completing the Supplement

Following are the supplement items and general instructions or definitions for each item as necessary. All of the items are not listed. As you read these instructions, you should refer to the items booklet (attachment) where all of the items and their answer categories are listed.

[**Note to Readers:** To conserve space, these instructions have been edited to include information only for those items used in classifying persons as displaced from a job (see Section C for details on the classification algorithm).]

**SD1.** During the last 3 calendar years, that is, January 1997 through December 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?

#### Purpose

The purpose of this question is to determine if a worker has lost a job involuntarily or left a job before it would have ended, in the last three calendar years. It is also used as a screening question to determine if the remainder of the "displaced workers" questions should be asked.

This question determines the job referred to throughout the displaced worker items (including the lost earnings items, SLE1-SLE22). Therefore, it is vital that you take into account the definitions of "lost job" and "involuntary separation" when you enter the answers to items SD1 through SD27.

#### **Definition of Lost Job**

Enter 1 (yes) in SD1 for persons who lost or left a job during the last three calendar years for the reasons stated in the question. Some workers will have lost more than one job in the last three calendar years. For these persons especially, you must clearly explain to the respondent that he/she should answer the displaced worker questions in terms of the lost job that was held the longest.

This would be the case even for persons currently unemployed because of a recent job loss. If they had previously (over the past three calendar years) lost a job which they had held longer than the job which they have recently lost, explain to them that the "displaced workers" questions refer to the earlier job. For example, if the person worked for Home Depot for 20 years, and lost that job in 1997, but found new employment in 1997 at Sears and subsequently lost that job, the reference job is the one at Home Depot.

#### **Definition of Involuntary Separation**

Enter 1 in SD1 if the person lost or left a job in the last three calendar years due to involuntary separation, as defined below:

1. Plant closed or moved - The place of business where the employee reported to work is no longer operating. The employer may have moved the business away or may have shut down the local operation permanently or temporarily. Include those persons that are offered relocation with an employer that moves, but turns down the offer.
2. Position or shift abolished - This could be caused by a company's losing a contract and terminating the jobs associated with that contract.
3. Insufficient work - Inadequate demand for a company's products or services, or for the individual's specific job.
4. Similar reasons - 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. If a person lost a job because his/her own business failed, enter 1. This would be true even for persons who are now operating another self-operated business, if the current business is different from the former one.

How to Complete

Enter 1 in SD1 if an individual retired because he was going to lose his/her job.

Enter 1 in SD1 if the worker was recalled by the same employer to do a different kind of work. For example, if the worker was formerly employed as a welder, but was recalled as an assembler, you should still enter 1 to report the loss of his job as a welder. However, enter 2 if the worker was recalled to the same job as a welder. Also, enter 2 if a person changed jobs with an employer with no period of layoff.

Enter 2 if the person left a job for personal reasons, such as going to school after a summer job or because of pregnancy. However, enter 1 if the worker chose to attend school after the plant closed permanently.

Enter 2 if the person was fired from a job because of poor work performance, disciplinary problems, or any other reason that is specific to that individual alone.

**SD2. Which of these specific reasons describes why you are no longer working at that job?**Definition of working "at that job".

Working "at that job" refers both to the specific employer and the kind of work done (i.e., a worker might have been laid off and rehired by the same employer in a different capacity. By the definition in Item SD1, that worker should still be reported as "displaced").

Only the reason that describes why the person is no longer at that job should be entered. For persons who were displaced from more than one job, "that job" should be the one that they held the longest.

How to Ask

Ask Item SD2 exactly as worded putting the emphasis on "at that job" and reading the list to the respondent. If the respondent indicated in Item SD1 that he or she has held and lost more than one job in the past three calendar years you might reword Item SD2 as follows: "For the job held longest, which of the following reasons describes why you are no longer working at that job?" Enter the precode for the main reason given.

<1> Plant or company closed down or moved.

If the employer closed the office or plant where the person worked, went out of business, moved out of the town or area and did not relocate workers (or workers did not want to relocate), or was acquired and did not keep the same workers, enter precode <1> for "Plant or company closed down or moved."

Plant or company operating but lost job because of:<2> insufficient work<3> position or shift abolished<4> seasonal job completed

Position or shift abolished could be caused by a company's losing the jobs associated with that contract.

Enter precodes <2-4> if the person lost his job and was rehired by the same employer but in a different capacity.

<5> Self-operated business failed

Enter precode <5> if a person closed his/her own place of business for reasons such as insufficient demand for their product or service or bankruptcy.

<6> Some other reason

Enter precode <6> for reasons not already covered.

**SD3. In what year did you last work at that job?**

Again, make sure that this answer refers to the job that was held for the longest period of time.

**SD4. Do you expect to be recalled to that job within the next 6 months?**

This question is asked of persons who reported losing their job in 1999. It is used to identify workers on layoff who expect to be recalled.

### **SD7 through SD15. Description of job formerly held.**

These items refer to the person's former job. Instructions for completing these items are the same as for completing basic CPS items IO1INT. (See instructions on pages B4-1 through B4-5 and C4-26 through C4-38 of your CPS Manual.)

**[Note to Readers:** Supplement item **SD7** reads as follows. Were you employed by government, a private company, or a non-profit organization, or were you self employed or working in a business owned by a member of your family?]

**SD18.** How long had you worked for that employer when that job ended?

#### Purpose

The purpose of this question is to find out how long the respondent had worked for his/her previous employer as described in Items SD10 or SD11. The words "that employer" will only appear if the respondent answered either <D> Don't know or <R> Refused in items SD10 or SD11. The instrument will fill with the response given in SD10 or SD11, if there is one. This will provide information on how long workers had been with the same employer before displacement.

#### Definition

A cumulative total of all spells of employment for a particular employer is what is being asked for this item. Breaks in service should be subtracted from the number of years reported, but there is no maximum time away from the job at which prior service is not counted. For example, John worked for J. C. Penney for 3 years; he took a year off to attend school then went back to J. C. PENNEY and worked for 2 years. He then lost that job. You would enter 5 years for this item.

## **Section B: CPS Concept of "Job"**

**Job.** A job exists when there is a definite arrangement for regular work every week, or every month, for pay or other compensation (e.g., profits, anticipated profits, or pay in kind, such as room and board).

A formal, definite arrangement with one or more employers to work on a continuing basis for a specified number of hours per week or days per month, but on an irregular schedule during the week or month, is also a job. (Bureau of the Census, 1999, p. B1-4. Additional details on this concept can be found in the CPS interviewing manual on pages B1-4 through B1-6).

## **Section C: Displaced Worker Classification Algorithm (2000)**

In 2000—and for the prior two administrations of the supplement as well (1996 and 1998)—a person (20 years of age or older) was classified as a displaced worker if the following conditions were met:

- (1) An entry of <1> ("yes") to **SD1**;
- (2) An entry of <1> (plant or company closed down or moved), <2> (insufficient work) or <3> (position or shift abolished) to **SD2**, but *excluding* those persons with entries of <2> or <3> to **SD2** who were separated from a job in the most recent year of the reference period (e.g., **SD3**=1999) AND who have an entry of <1> ("yes") to **SD4** (i.e., persons who have an expectation of return to work within six months);
- (3) An entry in **SD3** that fell within the specified three-year reference period (i.e., 1997, 1998, 1999);
- (4) A substantive entry to **SD7** (e.g., private company, government) *other than* self-employed (i.e., the latter are excluded from the displaced-worker designation); and
- (5) An entry in **SD18** of three years or longer. [Note: BLS also reports a larger estimate of displaced workers that ignores the previous tenure restriction (i.e., any substantive entry to SD18 would be acceptable).]