Some Thoughts on the Use of Field Tests to Evaluate Survey Questionnaires

James L. Esposito, *Bureau of Labor Statistics* June 2010 [1]

SECTION 1. Introduction

In contrast to other questionnaire-evaluation-methodology [QEM] papers to be presented at this workshop that focus either on specific evaluation methods (e.g., behavior coding; cognitive interviewing; experiments)—that involve recognizable if not standardized procedures—or on particular model-based approaches (e.g., latent class analysis; item response theory), this paper will focus on evaluation work that incorporates multiple evaluation methods and that by necessity is situated within field settings. *Field tests* are complex, resource-intensive, collaborative operations that draw upon the knowledge/information/data and skills possessed by various sources/agents (e.g., content and design specialists; interviewers and other field staff; respondents; statisticians) to optimize questionnaire design for the ultimate purpose of gathering high-quality data about a particular domain-of-interest. Because field tests represent evaluation work that occurs during specific phases of the *questionnaire-design-and-evaluation process,*[2] it will be necessary to preface the discussion of field-test methodology with a brief overview of the questionnaire-design-and-evaluation process (Section 2). This discussion will be limited to a cursory description of a *framework* that the author has found useful for situating field tests within the broader context of longitudinal (and potentially reiterative) design-and-evaluation work. With the framework as context, we then list and attempt to classify some of the methods and techniques that survey practitioners have at their

[1] The views expressed in this paper of those of the author and do not reflect the policies of the Bureau of Labor Statistics [USA].

[2] The term "questionnaire design-and-evaluation process" is intended as shorthand for a more inclusive process that incorporates questionnaire development work and also questionnaire redesign efforts.

Paper presented as part of the "Workshop on Questionnaire Evaluation Methods", 21-23 October 2009, at the National Center of Health Statistics, Hyattsville, MD. Final paper submitted June 2010.

disposal for evaluating questionnaires at various phases of the design-and-evaluation process (Section 3) and follow-up on that discussion with a very general introduction to field test methodology (Section 4). At that point, we move from the more abstract discussion of frameworks and methods to a more pragmatic discussion of field-test methodology in real-world settings (Section 5). In this section, we attempt to reconstruct what transpired in the course of an evaluation of a supplement questionnaire that actually involved a series of three separate field tests (conducted at two-year intervals), provide examples of method-generated qualitative and quantitative data, review how such data were analyzed and integrated, and offer some thoughts as to the utility of the various methods used. In the final section of the paper (Section 6), we offer some closing thoughts on the collaborative nature field-test methodology (Subsection 6.1) and attempt to address various issues relevant to future plans for incorporating field-test findings into the Q-Bank metadata structure (Subsection 6.2).

SECTION 2. Overview of the Questionnaire-Design-and-Evaluation Process

There are many excellent references in the survey methodology literature that describe various aspects of the questionnaire design-and-evaluation process (e.g., DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Eurostat, 2006; Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau, 2004; Lindström, Davidsson, Henningsson, Björnram, Marklund, Denell and Hoff (2004/2001); Martin, Hunter-Childs, DeMaio, Hill, Reiser, Gerber, Styles and Dillman, 2007; Platek, 1985; Snijkers, 2002), and these references provide detailed information about the various phases of the overall process, the work that would need to be accomplished at each phase, and the specialized knowledge that collaborators would need to possess to execute the process successfully.

To satisfy the modest objectives of this paper, the discussion will focus primarily on the measurement aspects of the process (as opposed to the parallel *representational* aspects of the process; see Groves et al. 2004, pp. 39-65, especially figures 2.2 and 2.5) and the perspective taken will be that of a survey practitioner with primarily questionnaire design-and-evaluation responsibilities.

To provide context for the discussion to follow, I will lean heavily on a heuristic framework developed by Esposito (2003, 2004a; 2004b)—a framework that has been constructed, in part, upon strong foundational ideas proposed by others (e.g., Belson, 1981; Cannell, Oksenberg , Kalton, Bischoping, and Fowler, 1989; DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Groves, 1987, 1989; Krosnick, 1991; Suchman and Jordan, 1990; Schaeffer 1991; Sudman and Bradburn, 1974, 1982; Thomas 1997; Tourangeau 1984; Turner and Martin, 1984; Willis, Royston and Bercini, 1991).

The framework (see **Table 1**) comprises two explicit dimensions and one implicit dimension: (1) Dimension One: eight design-and-evaluation phases (for both initial-design *and* redesign efforts); (2) Dimension Two: five sources of measurement error; and (3) the implicit dimension of time—coupled with the inevitability of social, cultural, and technological change.

With regard to the first dimension, *four core design phases* are specified:

- **P1:** *Observation.* The empirical foundation upon which "structures" of individual and integrated survey concepts/categories are built. Quality threats: Preconceived ideas/theories; limited field of observation.
- **P3:** *Conceptualization.* The process of simplifying/organizing domain-relevant observations into "structures" of individual and integrated survey concepts/categories. These "knowledge structures" represent the substantive elements that the design team

uses to develop questionnaire items and survey-relevant *metadata* (e.g., question objectives, conceptual definitions of key terms, interviewer training materials, classification algorithms). Quality threats: Preconceived ideas/theories.

- **P5:** *Operationalization.* The translation of domain-relevant concepts into questionnaire items and metadata. Quality threats: Inadequate design skills and/or metadata development.
- **P7:** *Administration.* Gathering self-report data by means of an interviewer-administered questionnaire.[3] Quality threats: Deficiencies associated with the various sources of measurement error and/or inadequate resources (time, staff and funding).

And four accompanying evaluation phases:

- **P2:** *Evaluation Work Targeting the Observation Phase*
- **P4:** *Evaluation Work Targeting the Conceptualization Phase*
- **P6:** *Evaluation Work Targeting the Operationalization Phase*
- **P8:** *Evaluation Work Targeting the Administration Phase*

With regard to the second dimension, five interdependent sources of measurement error are specified:

- **S1:** *Questionnaire Design-and-Evaluation Team: Content Specialist.* Content specialists are individuals who possess subject-matter expertise (e.g., survey sponsors; program managers) with regard to a particular domain-of-interest (e.g., health; labor-force dynamics; income and wealth; demographics).
- **S2:** *Questionnaire Design-and-Evaluation Team: Design Specialist.* Design specialists are typically survey practitioners who, in collaboration with content specialists, design and evaluate questionnaires; they also assist in the development of ancillary metadata, like interviewing manuals and classification algorithms.

[3] The development of the current framework reflects the author's research experiences with interviewer-administered surveys primarily, but modifying the framework to encompass other types of surveys (e.g., self-administered) would not be difficult.

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

			INTERDEPENDENT SOURCES OF MEASUREMENT ERROR **(at P7 or RP7)**				
			Questionnaire D-and-E Team		**Information/Data Collection Context**		
		REDESIGN	*Content Specialist (1)*	*Design Specialist (2)*	*Interviewer (3)*	*Respondent (4)*	*Mode (5)*
	RP8	Evaluation	C_{R81}	C_{R82}	C_{R83}	C_{R84}	C_{R85}
	RP7	***Administration***	$\mathbf{C_{R71}}$	$\mathbf{C_{R72}}$	$\mathbf{C_{R73}}$	$\mathbf{C_{R74}}$	$\mathbf{C_{R75}}$
Questionnaire	RP6	Evaluation	C_{R61}	C_{R62}	C_{R63}	C_{R64}	C_{R65}
***Redesign* and**	RP5	*Operationalization*	C_{R51}	C_{R52}	C_{R53}	C_{R54}	C_{R55}
Evaluation	RP4	Evaluation	C_{R41}	C_{R42}	C_{R43}	C_{R44}	-
Phases	RP3	*Conceptualization*	C_{R31}	C_{R32}	C_{R33}	C_{R34}	-
	RP2	Evaluation	C_{R21}	C_{R22}	C_{R23}	C_{R24}	-
	RP1	*Observation*	C_{R11}	C_{R12}	C_{R13}	C_{R14}	-
		INITIAL DESIGN	*Content Specialist (1)*	*Design Specialist (2)*	*Interviewer (3)*	*Respondent (4)*	*Mode (5)*
	P8	Evaluation	C_{81}	C_{82}	C_{83}	C_{84}	C_{85}
	P7	***Administration***	$\mathbf{C_{71}}$	$\mathbf{C_{72}}$	$\mathbf{C_{73}}$	$\mathbf{C_{74}}$	$\mathbf{C_{75}}$
Questionnaire	P6	Evaluation	C_{61}	C_{62}	C_{63}	C_{64}	C_{65}
***Design* and**	P5	*Operationalization*	C_{51}	C_{52}	C_{53}	C_{54}	C_{55}
Evaluation	P4	Evaluation	C_{41}	C_{42}	C_{43}	C_{44}	-
Phases	P3	*Conceptualization*	C_{31}	C_{32}	C_{33}	C_{34}	-
	P2	Evaluation	C_{21}	C_{22}	C_{23}	C_{24}	-
	P1	*Observation*	C_{11}	C_{12}	C_{13}	C_{14}	-

Observational base: The **domain-of-interest** as embedded in a "reality" of ceaseless activity (behavior and events) and of durable-yet-mutable relationships (some real, some spurious)—a world within which the observer is an active participant.

Paper presented as part of the "Workshop on Questionnaire Evaluation Methods", 21-23 October 2009, at the National Center of Health Statistics, Hyattsville, MD. Final paper submitted June 2010.

- **S3:** *Interviewer*. Interviewers are members of a field organization that receive special training in their primary role as data collectors—and, as such, are expected to serve in this role for both production surveys *and* evaluation studies.
- **S4:** *Respondent*. Respondents are data providers that have been selected from a larger population of individuals about whom a survey sponsor wishes to gather specific information about the domain-of-interest. To be informative, evaluation studies (especially field studies) should draw samples from the same population of individuals as that to be used (or currently in use) for the production survey.
- **S5:** *Mode*. The term *mode* refers to the various technical methods/procedures by which means survey organizations gather data bout the domain-of-interest (e.g., telephone surveys via paper-and-pencil questionnaire or via centralized computer-assisted interviewing; face-to-face interviewing; via paper-and-pencil questionnaire or via centralized computer-assisted interviewing; self-administered interviewing via a paper questionnaire or a computerized instrument).

Several additional aspects of the framework are worthy of note (Esposito 2003, p. 55, with modifications):

- First, it is presumed that design-and-evaluation work can and often does overlap across phases and that movement between certain phases (P1 through P6) is bidirectional and potentially iterative.

- Second, the phrase "interdependent sources of measurement error" has been adopted to reflect the view that measurement error is presumed to be the outcome of collaborative/interactive processes involving the various sources of error identified in Table 1. Within a given data-collection context, measurement error is presumed to be a byproduct of role- and task-specific activities—Sudman and Bradburn's (1974) terminology (cf. Platek 1985)—that manifest themselves during the survey administrative phase (P7 or RP7). Various role- and task-specific activities that are performed inadequately at prior design-and-evaluation phases (P1 through P6) can be viewed as *precursors* to measurement error.

- Third, the actual performance of role- and task-specific activities—represented as generically-labeled cell entries (e.g., C_{12})—is presumed to vary across questionnaire design-and-evaluation efforts. Whether or not a particular cell has an entry would depend on whether specific cell-related activities were conducted. For example, if content specialists are not involved in pretesting work conducted during the initial questionnaire design, then cell C_{61} would be left blank (i.e., signifying no record of collaborative activity during P6). Empty cells are problematic in that they represent deficits in knowledge/information/data that have the potential to affect/increase the magnitude of measurement error realized during an ***Administration phase*** (e.g., P7) and assessed during the appropriate evaluation phase (e.g., P8).

- And lastly, social, cultural and technological change also plays a crucial role in the measurement process. Unless continuously monitored and accounted for by content and design specialists, rapid change within a given domain-of-interest can have a substantial effect on measurement error.

SECTION 3. Questionnaire Evaluation Methods

When it comes to evaluating questionnaires, survey practitioners can draw upon a broad array of analytical methods (and techniques).[4] Some of these methods generate qualitative data primarily, some quantitative data, and some methods yield both types of data. Some methods appear more useful for evaluating interviewer and/or respondent performance (e.g., paradata analysis), some more useful for evaluating the performance of content and/or design specialists (i.e., with specific regard to the "performance" of specific questionnaire items), and some methods appear useful for evaluating all of the

[4] These two terms, *methods* and *techniques*, appear to be used interchangeably in the literature. In some contexts, the former can reasonably be viewed as primary or dominant, and the latter as secondary or subordinate, such as when one refers the use of retrospective probes (a specific procedural technique) when conducting cognitive interviews (the host method).

above (e.g., behavior coding). Some methods (e.g., focus groups) are broadly applicable in that they can be used productively during any of the evaluation phases identified in Table 1 (i.e., P2, P4, P6 and P8), while others seem best employed during a particular phase (e.g., the method of reinterview, at P8). Elaborating on the latter generalization, there seems to be a gathering consensus in the literature that: (1) given the properties associated with specific methods, there may be an optimal sequence for using specific classes of methods across evaluation phases (see **Table 2**); and (2) the use of multiple methods at any particular evaluation phase—given strengths and weaknesses associated with all methods—reduces the risk of misidentifying potentially serious design flaws.

When conducting field studies (e.g., at P8), practitioners make important decisions not only with respect to selecting which particular methods to utilize in evaluating an existing production questionnaire, but also in terms of how to sequence the methods to optimize their utility. For example, sometimes a method applied/conducted early in the sequence of a particular multiple-method evaluation effort can be useful in making enhancements to procedures for gathering other evaluation data subsequently using a second method, like when the coding of interviewer-respondent interactions (behavior coding) during questionnaire administration yields insights regarding a key questionnaire item that prompt a researcher to add one or more unscripted probes to the protocol of a focus group used to debrief interviewers soon thereafter.

Table 2. A Non-exhaustive List of Questionnaire Evaluation Methods

Evaluation Methods	Phase(s)	Locus of data collection	Comments
Anthropological/Ethnographic Methods			
▪ Unstructured interviews	P4	Lab, office or field settings	
▪ Unobtrusive observation	P2	Field settings	For example, Webb et al., 1966.
▪ Participant observation	P2	Field settings	
▪ Comparative analysis	P2, P4	Field settings	For example, Glaser and Strauss, 1967/1999.
▪ Rapid Assessment Process	P2, P4	Field settings	See Beebe 2001.
Cognitive Methods			
▪ Intensive interviews	P4, P6	Lab, office or field settings	A precursor of the modern, post-CASM method of cognitive interviewing (e.g., see Royston 1989; Willis 2005).
▪ Cognitive interviews	P4, P6	Lab or office	Variations: Concurrent vs. retrospective think-aloud interviews, possibly incorporating other techniques (see Willis 2005).
▪ Ancillary cognitive techniques	P4, P6	Lab or office	Examples: Confidence ratings; paraphrasing; free and dimensional sorts; response latency; scripted and unscripted probes; memory cues.
Expert Review Methods			
▪ Expert panels	P4, P6	Lab, office or field settings	
▪ Questionnaire appraisal systems	P4, P6	Office	For example, Lessler and Forsyth, 1996; Willis and Lessler, 1999.
Debriefing Methods			
▪ Post-interview follow-up probes (and/or vignettes)	P6, P8	Field settings	For example, Martin 2004.
▪ Post-interview follow-up structured interviews	P6, P8	Field settings	For example, Belson 1981; Sykes and Morton-Williams, 1987
▪ Calendar method	P6, P8	Lab, office or field settings	Potentially useful in evaluating questionnaire data collected during field tests (e.g., see Belli, Lee, Stafford and Chou, 2004).
▪ Ad hoc debriefing questionnaires	P6, P8	Office or field settings	Focus: Interviewers (in most cases).
▪ Interview logs	P6, P8	Field settings	Useful as documentation when debriefing interviewers.
▪ Rating scales	P6, P8	Office or field settings	Useful when embedded in debriefing sessions with interviewers.
▪ Focus groups	P4, P6, P8	Lab, office or field settings	A general method that can be used to gather information from any participant group (e.g., informants, interviewers, respondents, practitioners, sponsors).

Interaction-Coding Methods			
▪ Behavior coding	P6, P8	Field settings	For a comprehensive review, see Ongena and Dijkstra, 2006.
▪ Conversation analysis	P6	Lab or field settings	For example, Schaeffer 2002.
Field Test Methods			
▪ Pilot tests (survey simulations)	P6	Field settings	Focus: Draft questionnaire. Potentially resource intensive.
▪ Production-survey tests	P8	Field settings	Focus: Existing/production questionnaire. Resource intensive.
▪ Reinterviews	P8	Field settings	Focus: Existing/production questionnaire. Resource intensive.
▪ Experiments	P6, P8	Field settings	Focus: Draft or existing/production questionnaire. Resource intensive. For example, Tourangeau 2004.
▪ Split-sample/panel tests	P6, P8	Field settings	Focus: Draft or existing/production questionnaire. Resource intensive. For example, Fowler 2004.
Statistical Analyses and Modeling			
▪ Response-distribution analysis	P6, P8	Field settings	
▪ Nonresponse analysis	P6, P8	Field settings	For example, see Groves and Peytcheva (2008), Groves (2006) and Olsen 2006.
▪ Latent class analysis	P6, P8	Field settings	For example, Biemer 2004
▪ Item response theory modeling	P6, P8	Field settings	For example, Reeve and Mâsse, 2004.
Computer-linked Methods			
▪ Paradata analysis	P6, P8	Field settings	

Primary sources: DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Esposito and Rothgeb, 1997; Forsyth and Lessler, 1991; Jobe and Mingay, 1989; Royston, Bercini, Sirken and Mingay, 1986; Willis 2005.

SECTION 4. Field Test Methodology: A Brief Introduction

Field tests, which for evaluation purposes involve the actual administration of either a proposed survey questionnaire (i.e., P6 field test) or a ongoing production survey questionnaire (i.e., P8 field test), come in a "various colors and sizes" from large-scale, multiple-method, multiple-phase undertakings, like the redesign of the Current Population Survey [**CPS**] (e.g., Esposito and Rothgeb, 1997) to small-scale, rapid-turn-around pilot tests of questionnaires that gather data on a specific topic (e.g., cell-phone usage in the United States; see Tucker, Brick and Meekins, 2007; Esposito 2005). The CPS redesign, for example, involved three phases—that is, three separate field tests conducted over a four-year period (1990 through 1993)—and made use of the following evaluation methods (and techniques):

- Split-panel tests—making use of both response-distribution and item-nonresponse analyses
- Interviewer debriefings—making use of both focus groups and structured questionnaires
- Behavior coding
- Respondent debriefing—making use of both vignettes and post-interview follow-up probes

Not all surveys can command the resources that were required to redesign the CPS, one of two principal labor force surveys conducted in the United States each month; and it is difficult to imagine how informed design decisions could have been made (e.g., question content and sequencing) in the absence of the data generated during these three evaluation phases (see **Appendix,** Tables A-1 and A-2). Yet while few American national statistical surveys can match the importance of the CPS in terms of measuring key aspects of the U.S. economy, there are many other national statistical surveys that gather important data, that need to be carefully evaluated during initial design (and periodically thereafter), and that require substantial resources and a viable research plan to accomplish successfully a variety of evaluation tasks. It is to a brief

discussion of the latter two elements—available resources (subsection 4.1.) and a viable research plan (subsection 4.2.)—that we now turn.

Subsection 4.1. *Resources Required to Support Effective Questionnaire Design-and-Evaluation Research Efforts.* We live and work in a world of limited resources and such limitations constrain our ability to gather information that would be useful in optimizing the questionnaire-design-and-evaluation process. The following represents a short list of critical resources and some of the constraints often associated with suboptimal resource allocation:

- DOMAIN-RELEVANT KNOWLEDGE/INFORMATION/DATA: *The relevant "who, what, when, where, how and why" associated with the domain-of-interest.* Insofar as domain-relevant knowledge/information/data provide the foundations upon which to plan and carry out constructive design-and-evaluation work, these resources are most useful when grounded in first-hand observations of real-world behavior and events. In addition to the substantial time and money required to amass relevant knowledge/information/data, content specialists are often constrained by their conceptual models regarding the domain-of-interest (e.g., the essential nature of mental and physical disabilities) and/or by official "definitions" of that domain—which in some cases may differ substantially from the views and models of other subject-matter experts (e.g., those advocating person- vs. context-based models of disability) and from the myriad lay perspectives that members of general public have constructed on the basis of their experiences.
- STAFF: *The professionals available to make contributions to the process (e.g., content specialists; design-and-evaluation specialists; programmers/authors; operations/production managers).* The success of any challenging design-and-evaluation effort will be compromised to the extent that available staff are limited and/or inexperienced in the roles they are expected to perform. For example, if the practitioners responsible for conducting evaluation research are only familiar with one or two evaluations techniques (e.g., say focus groups and/or behavior coding), that greatly limits the type and amount of analytical information/data available to survey sponsors with which to make decisions regarding content changes and design modifications.

- TIME AND FUNDING: *The amount of time and money required to support and execute the various phases of the design-and-evaluation process*. The process of designing and evaluating surveys used to gather important social and economic data—about such topics as: poverty, health and safety, energy use, consumer prices, employment and industrial activity—constitutes an expensive and time-consuming undertaking; and when juxtaposed with other national priorities, such endeavors must compete for scarce taxpayer dollars. Given the critical importance such surveys and the investments required to ensure that the data produced are both reliable and valid, it is essential that national statistical organizations have access to the resources needed to conduct this work in an efficient and highly professional manner, and that these organizations be held accountable for the quality of the data they provide.

We might note here that, in some cases, a lack of resources in one area (e.g., available time to complete work) can be offset by the availability of more-than-sufficient resources in other areas (e.g., funding and available staff). For example, in a situation where the time available to field test a given draft questionnaire is temporally constrained, but project-available funds and staff are more than sufficient to complete a multiple-method evaluation phase in a timely fashion, a management decision can be made to reassign staff from projects that are less time-sensitive to the one that is highly time sensitive. But there are limits to resource substitution. For example, if the knowledge/information/data made available by a survey sponsor for questionnaire-design purposes (e.g., concepts and measures associated with, say, the measurement of mental and physical disabilities) are not consistent with what most observers would accept as representative or valid regarding the domain-of-interest (e.g., *their* understanding of mental and physical disabilities), but the sponsor insists that a draft questionnaire be developed and evaluated in a short span of time using only those elements provided, then there appears to be little a survey organization (or its involved practitioners) can do other than to document their misgivings with respect to the specifications provided and/or possibly move to rescind their contract with the

sponsor—if such an action is feasible, politically and contractually. Better to lose a contract than to compromise the organization's standards for gathering high-quality data.

Subsection 4.2. *Developing a Timeline and a Viable Research Plan for Conducting Field Tests.* What constitutes a reasonable timeline and a viable research plan for conducting a specific field test depends greatly on the scope of the work that needs to be accomplished, available resources, and a survey practitioner's prior experience with such tests (and with various evaluation methods). Generally speaking, the more important the survey, the greater the scope of work and the more likely it is that (marginally) sufficient resources will be made available for conducting a field test. The generalized timeline and research plan described below has worked for the present author in evaluating various supplements to the Current Population Survey, but may not be optimal for survey practitioners conducting field tests under different conditions and constraints. It has been my experience that the most satisfying evaluation efforts are those that gather data/information from various sources (e.g., survey sponsors; respondents; interviewers, design specialists)—and when we have been successful at doing so, it has usually required adoption of a multiple-method evaluation strategy. This particular evaluation strategy, which represents an idealized composite of a number of prior field tests, is summarized in **Table 3** and discussed very briefly below.

Table 3. A Generic Timeline of Research Activities for Conducting a Small, Phase-Eight [P8] Field Test

Lead/Lag Time Relative to Onset of Field Test [P7]	**Research Tasks: Performed by Survey Practitioner(s) with Other Members of Evaluation Team**
Lead time: 8 to 12 months	▪ Confer with the survey sponsor as to: (i) the scope of work; and (ii) the timeframe for completing the work. ▪ Determine what level of resources (i.e., staff and funding) will be needed to conduct the evaluation work within the timeframe specified. ▪ Request objectives and specifications for current set of questionnaire items. ▪ Review conceptual issues with the sponsor and, as needed, conduct a literature review to enhance knowledge and understanding of the domain-of-interest. ▪ Review prior evaluation research and, time permitting, conduct additional evaluation work as needed (e.g., expert panels; expert review of existing production questionnaire; cognitive interviews). ▪ Alert the Office of Management and Budget [OMB] as to evaluation plans and obtain necessary clearances to proceed.
Lead time: 8 to 10 months	▪ In collaboration with survey sponsor (and/or other content specialists), develop a set a post-interview follow-up probes to evaluate critical items on the existing questionnaire. ▪ Commence work on developing interviewer instructions for administering debriefing probes to respondents. ▪ Confer with Field Operations representatives to review the proposed research plan (and continue to do so on an as-needed basis).
Lead time: 4 to 6 months	▪ Submit specifications for debriefing probes to programming authors. ▪ Commence work on a focus-group protocol for debriefing interviewers.
Lead time: 2 to 3 months	▪ Test follow-up probes that have been designed for debriefing respondents. ▪ Train survey practitioners on procedures for conducting behavior coding and focus groups (as needed). ▪ Design (or modify the existing) log materials to be used by interviewers to record problems experienced when administering key questionnaire items. ▪ Design (or modify the existing) rating form to be used by interviewers in independently assess the severity of problems encountered when administering items on the existing production questionnaire. ▪ Design (or modify the existing) behavior-coding form to be used for coding interactions between interviewers and respondents during questionnaire ***Administration*** (P7)
Lead time: 7 to 10 days	▪ Send interviewer logs (and all supporting documentation) to managers at centralized CATI locations for subsequent distribution to participating interviewers

During ***Administration*** (P7) or soon thereafter	▪ All data-collection days: Interviewers (i.e., those selected to be focus group participants) record problems experienced when administering key questionnaire items. Interviewers administer follow-up probe questions to respondents. ▪ Days 1 through 3: Conduct behavior coding at centralized CATI location(s). ▪ Days 4 and 6: At various centralized locations, conduct (and audiotape) focus groups with pre-selected interviewers using a rating scale designed to assess the magnitude of problems.
Lag time: 1 to 3 months	▪ Create a behavior-coding database and enter data. Prepare draft report for review and comment. ▪ Transcribe focus-group audiotapes, organize comments and compute means and standard deviations from ratings data provided by interviewers. Prepare draft report.
Lag time: 4 to 6 months	▪ Obtain and review respondent debriefing data (i.e., post-interview follow-up probes) from data-collection agency. Conduct statistical analyses (as needed). Prepare draft report for review and comment.
Lag time: 6 to 10 months	▪ Prepare a "composite report" that integrates findings from the various evaluation methods employed and distribute draft report to survey sponsor for review and comment. Review sponsor comments, make necessary modifications, finalize and submit report to sponsor and other interested parties.

As can be seen inferred from the content of Table 3, field tests are resource-intensive undertakings (i.e., time, money and staff); and the tasks that must be completed during the actual evaluation phase (P8, in this example) depends to a large extent on the scope and quality of work that has been undertaken and accomplished in earlier phases of the design-and-evaluation process (P1 through P7). Field tests require a thorough understanding of both survey content and evaluation methodology. Accumulating knowledge/information/data about survey content and prior evaluation work occurs early in the timeline; the tasks of developing and distributing instructional and evaluation materials follow. These early, more time-consuming tasks are often followed by a relatively intense period of data collection using a variety of evaluation methods applied in a predetermined sequence—more about this later. The collection of evaluation data (at P8) is followed by a more deliberate period of data review-and-consolidation, analysis and report writing. Although the overall process is necessarily collaborative, there are times during which the survey practitioner is working as an independent agent; to perform well, one needs to learn to be competent and comfortable with both aspects of the practitioner's role.

SECTION 5. A Case Study of Field Testing in Practice: The Displaced Worker Survey [5]

It is one thing to talk about the questionnaire design-and-evaluation process in the abstract and quite another to describe and summarize the sometimes frenetic activity of an actual field test. Retrospective summaries of such work may be viewed—generously, in my view—as relatively benign reconstructions/understatements of what actually took place during the "heat of battle." That said, and in an effort to provide some additional detail and data on the actual implementation of field-test methods and procedures, we now turn to a discussion of a series of

[5] In this section of the paper, the author draws heavily on a article published previously in the Journal of Official Statistics (Esposito 2004).

field tests designed to evaluate a supplement to the Current Population Survey [**CPS**] that gathers data on worker displacement.

Subsection 5.1. *Background.* In the early 1980s, the American economy was staggered by two recessions that were especially hard on manufacturing industries, particularly steel and automobile production. In an effort to assess the effects of these developments on the labor force, a small group of labor economists (content specialists) at the Bureau of Labor Statistics, in collaboration with design specialists at the Census Bureau, set about to design 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 CPS in 1984. Although the DWS was intended to be a *one-time* survey, the data generated had utility for both internal and external users and, as a result, has been administered biennially ever since. The primary objective of the supplement is to estimate the number of workers who have lost or left a job for specified displacement reasons and to collect data on the types of jobs that these workers have lost or left.

In June 1995, a survey practitioner (i.e., the present author) was asked to review the DWS to identify potential sources of measurement error. This "expert" review identified a number of potential problems with the DWS questionnaire: (1) problematic question wording, especially with respect to two key supplement items used to classify target persons as displaced or not displaced from a job; (2) ambiguous conceptual terminology; and (3) unclear or incomplete question specifications. Concern about these potentially problematic issues prompted the supplement sponsors to authorize and fund a small (P8) field test, and this test was conducted in February 1996. [**Note:** Almost all of the evaluation data to be reviewed herein focuses on supplement items SD1 and SD2 (see **Table 4**). The reason for focusing on these two items is

that they carry most of burden for classifying workers who have separated from jobs during the reference period as displaced or not displaced.]

Table 4. Supplement Items SD1 and SD2 (Adults, Unweighted Data, 1996—2000)

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

Perhaps the most crucial aspect of any evaluation effort—including field tests—is gathering survey ***metadata***, such as question objectives, conceptual definitions of key terms, interviewer training materials, classification algorithms, prior evaluation reports, and information regarding changes to question concepts, wording and/or sequencing over the lifecycle of the survey. [6] This can be a challenging exercise, especially if the survey questionnaire has reached an "advanced age" (e.g., first administered circa 1984 or earlier). Given that arbitrary definition, which

[6] Why is such metadata crucial? Because, in principal, the delineation of survey-relevant observational and operational details and the specification of concepts supporting question/questionnaire design provide the *basis* for conducting evaluations (P2, P4, P6 and P8) of any one (or all) of the four core phases (P1, P3, P5 and P7) of the questionnaire design-and-evaluation process. See Dippo and Sundgren (2000) for a gentle introduction to metadata.

coincides with the advent of CASM (i.e., the copyright date of the seminal monograph on cognitive aspects of survey methodology), the DWS questionnaire qualifies; but fortunately, locating useful metadata (e.g., interviewer memoranda describing various concepts and data-collection procedures; an article in a government publication that reported findings from the first administration of the DWS) was not particularly difficult, though apparently there was not a lot of metadata to be found. With access to such reference materials, it was possible to get a sense of displaced worker concept and the manner in which the sponsor intended to measure this concept (see **Table 5**).

Subsection 5.2. *Field Test Methodology and Selected Findings [SD1 and SD2]*. The research conducted on the DWS during the period 1995-2000 is based on a multiple-method approach to questionnaire evaluation that was used in the early 1990s by researchers at the BLS and the Census Bureau to redesign the CPS (e.g., see Esposito and Rothgeb, 1997). Various research methods are used to gather qualitative and quantitative data about different aspects of the survey measurement process (e.g., the interpretation of key concepts; the comprehension of question meaning; the efficiency of interviewer-respondent interactions). Data gleaned from multiple methods can be compared and contrasted to provide researchers with a more comprehensive picture of how well target questions are meeting their stated objectives (e.g., Cannell et al., 1989; Oksenberg, Cannell and Kalton, 1991; Sykes and Morton-Williams, 1987).

Table 5. Relevant Metadata Associated with the Displaced Worker Concept and Two Key Displaced Worker Questions, SD1 and SD2 (1998 Field Test)

A Working Definition of Displaced Worker	"While there has never been a precise definition for [displaced workers], the term is generally applied to persons who have lost jobs in which they had a considerable investment in terms of tenure and skill development and for whom the prospects of reemployment in similar jobs are rather dim … (Flaim and Sehgal, 1985, p.4)."
SD1**	**During the last 3 calendar years, that is January 1995 through December 1997, 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. *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. … *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: 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. Position or shift abolished - This could be caused by a company's losing a contract and terminating the jobs associated with that contract. Insufficient work - Inadequate demand for a company's products or services, or for the individual's specific job. 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. [**Note:** Table 5 continues on the next page.]

Table 5 (continued)

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.** **<1> Plant or company closed down or moved** **Plant or company still operating but lost or left job because of:** **<2> Insufficient work** **<3> Position or shift abolished** **<4> Seasonal job completed** **<5> Self-operated business failed** **<6> Some other reason**
	Question Objective: To determine the specific reason for job loss with the understanding that if more than one job was lost during the reference period, the respondent would be instructed to report on the longest-held job. [Note: Added by author. Not included in Census Bureau memorandum to interviewers (February 1998).]
	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* [emphasis added] 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.
Classification Algorithm	With one exception (i.e., persons displaced/laid-off from a job in the most recent year of the reference period who are expecting to be recalled to that job), persons categorized into one of the first three response options of SD2 are classified as displaced workers. Those whose answers are coded into one of the latter three response options (4 through 6) are asked no further DWS questions and are *not* classified as displaced workers.
**	**Note:** Almost all of the conceptual and procedural metadata for supplement items SD1 and SD2 was retrieved from the CPS Field Representative and CATI Interviewer Memoranda for the Displaced Worker Supplement, Number 1998-02 (Bureau of the Census, February 1998).

Subsection 5.2. *Field Test Methodology and Selected Findings [SD1 and SD2].* The research conducted on the DWS during the period 1995-2000 is based on a multiple-method approach to questionnaire evaluation that was used in the early 1990s by researchers at the BLS and the Census Bureau to redesign the CPS (e.g., see Esposito and Rothgeb, 1997). Various research methods are used to gather qualitative and quantitative data about different aspects of the survey measurement process (e.g., the interpretation of key concepts; the comprehension of question meaning; the efficiency of interviewer-respondent interactions). Data gleaned from multiple methods can be compared and contrasted to provide researchers with a more comprehensive picture of how well target questions are meeting their stated objectives (e.g., Cannell et al., 1989; Oksenberg, Cannell and Kalton, 1991; Sykes and Morton-Williams, 1987).

Three principal evaluation methods were used during each phase of this multiple-phase research effort: (1) interviewer debriefings; (2) behavior coding; and (3) respondent debriefings. The rationale for the repeated use of these three methods is as follows. First, collectively, the three general methods capture or reveal the perspectives of the various parties involved in the survey measurement process—interviewers, respondent, content and design specialists. Second, the survey practitioner responsible for this evaluation work had used these methods in prior research and he had found them to be efficient, effective and relatively inexpensive to employ. And third, to maintain a level of methodological comparability across phases, we wanted the replications to be as uniform as possible.

Interviewer Debriefing. Focus groups were used as the principal method for gathering evaluation information from interviewers (e.g., DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Morgan, 1988). During the phase-two evaluation, we also incorporated a rating form with a target-question *rating scale* (see Table 8, bottom). 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. Several days prior to administering the DWS, interviewers selected to participate in the focus groups were given *log forms* (see **Appendix**, Table A-3) 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—SD1 and SD2, specifically, and, in phase three, respondent debriefing items. An extensive protocol of scripted probe questions was used to guide the group discussion and stimulate interviewer feedback (e.g., see **Table 6**). Focus group sessions were audiotaped and written summaries were prepared from these tapes. Some examples of the qualitative and quantitative data obtained from interviewers debriefings can be found in **Tables 7 and 8**, respectively.

Table 6. Examples of Scripted Interviewer Debriefing Questions (1998 Field Test)

SD1	▪ Did you experience any difficulty reading this question in its entirety before respondents provided an answer? ▪ Did any respondents appear to have difficulty understanding the phrase: "lose a job or leave one"? ▪ Did respondents appear to understand the meanings of the various displacement conditions provided in the body of this question? If not, what types of problems did they seem to have? ▪ How clear were interviewer's instructions in providing descriptions of the various displacement conditions? ▪ Was the phrase "or another similar reason" causing any problems for respondents? ▪ Did respondents ask about any situations other than the displacement conditions specifically mentioned in this question? If yes, what types of situations did they mention?
SD2	▪ Did you have difficulty reading this question in its entirety? ▪ Were you able to read all 6 response options without being interrupted by the respondent? ▪ 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"? What types of responses did you categorize as "some other reason? ▪ How frequently did you read the READ AS NECESSARY statement? ▪ Was the READ AS NECESSARY statement confusing to respondents? ▪ Were respondents able to recall which job they held the longest? ▪ Were there confusions about identifying a particular job which would serve as the focus of later questions? ▪ Did respondents understand the meaning of each of the displacement reasons provided in the question? If not, which reasons did respondents fail to understand?

Table 7. A Sampling of Qualitative Data Generated from the Focus Groups Conducted with CPS/DWS Interviewers (1998 Field Test)

SD1

- Interviewers in all three focus groups mentioned problems that various respondents had experienced in interpreting the intent of this question. Much of this confusion appeared to center on the meaning of the phrase "or another similar reason". Given their answers to subsequent questions (i.e., SD2), some respondents clearly interpreted the question more broadly than intended—to include jobs that may have been lost or left during the reference period for *any reason* (e.g., to take a better job; to go back to school; to start a business). Interviewers themselves were not completely sure what this phrase encompassed. Most apparently assumed that it meant for a reason *similar to* one of the reasons that was specifically mentioned in the body of the question. In fact, when interviewers realized (at SD2) that the reason was not similar (e.g., the target person had left a job voluntarily to take something better with a different employer), some interviewers felt obligated to skip back to SD1 and change the entry from "yes" to "no".

SD2

- Several interviewers found the wording and format of SD2 to be awkward. And it did not help to have the "read-if-necessary" statement embedded between the first part of the question and the to-be-read response options. With regard to coding responses, one interviewer offered the following insightful comment: "There are lots of times when you ask somebody [this sort of question], and then they tell you something and *you decide* what category it goes into… (italics added)." Along similar lines, a second interviewer said that, even though the question asked for a specific reason, [some] respondents would "give you everything—they would tell you something that wasn't listed and you would have to read the question over again."
- Most interviewers apparently did not read the "read as necessary" statement unless respondents volunteered information (in their responses to SD1) that the target person had lost more than one job during the reference period.

Table 8. Interviewer Ratings for Supplement Items SD1 and SD2 (1998 Field Test)

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

Note: Interviewers were asked to rate problematic supplement items using the following 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)*

Behavior Coding. Behavior coding was used as the principal method for gathering evaluation information about the interviewer-respondent interactions during supplement administration (see Ongena and Dijkstra, 2006, for an extensive review). The method of behavior coding involves a set of procedures which have been found useful in identifying problematic questionnaire items (e.g., 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). 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—the interviewer side or the respondent side—a maximum of two behavior codes could be assigned (e.g., "AA" and "INT" if, for example, the respondent interrupted the interviewer with an adequate answer before the interviewer could finish asking the question); however, for most exchanges, only one interviewer code and only one respondent code was assigned. Extended interactions were coded, when possible, for SD1 and SD2 (and for other key supplement items). **Table 9** provides a sampling of behavior coding data (for selected problem codes only) for all three filed tests.

Table 9. Behavior Coding Data for Selected Items (1996, 1998 and 2000 Field Tests)

Field Test	Item(s)	Interviewer Codes		Respondent Codes			
		E	MC	AA	IA	RC	INT
1996	SD1	65% (33/51)	16% (8/51)	88% (42/48)	2% (1/48)	8% (4/48)	19% (9/48)
	SD2	29% (2/7)	57% (4/7)	67% (4/6)	33% (2/6)	0% -	17% (1/6)
1998	SD1	71% (96/135)	13% (18/135)	88% (119/135)	10% (13/135)	1% (1/135)	25% (34/135)
	SD2	0% -	72% (13/18)	56% (10/18)	28% (5/18)	0% -	39% (7/18)
2000	SD1	69% (82/119)	18% (22/119)	93% (110/118)	5% (6/118)	0% -	13% (15/118)
	SD2	29% (4/14)	43% (6/14)	60% (6/10)	40% (4/10)	0% -	0% -

Notes. Data are presented for two key supplement questions (SD1 and SD2) 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).

Respondent Debriefing. We used *post-interview follow-up probes* as the principal method for gathering information from respondents as to how well survey concepts were being understood (e.g., Campanelli, Martin and Creighton, 1989; Campanelli, Martin and Rothgeb, 1991; Hess and Singer 1995; Martin 2004; Oksenberg, Cannell and Kalton, 1991; cf. Schuman, 1966). A small interdisciplinary team of design and content specialists drafted the respondent debriefing questionnaire. The total number of debriefing questions varied from one phase to the next. The debriefing items were designed: (1) to gather job-related information that was relevant to job separation concepts, and (2) to determine whether item-specific problems existed that might jeopardize an accurate count of displaced workers. Each debriefing question was designed with a specific objective in mind (see **Table 10** for several examples). Answers to debriefing

questions were very useful in helping the research team to detect potential sources of measurement error (see **Tables 11A through 11D**). 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.

Table 10. Examples of Respondent Debriefing Items (1998 Field Test)

SDB1	Earlier you told me that me that you had lost or left a job in the past three calendar years [*fill with displacement reason from SD2*]. Did you lose that job or did you leave that job?
	Rationale: The supplement sponsor wished to know what percentage of displaced workers had lost a job relative to those who had left a job. We presumed the respondent could make this distinction without guidance from the sponsor. This probe also is used to channel *job leavers* to specific follow-up probes.
SDB3Z	Did you ever return to work for that employer, for even a short period of time?
	Rationale: For persons reported to have *lost, left, or retired from* a job during the reference period for a displacement reason, to determine if the person returned to work for that employer, even briefly. This item is an attempt to identify individuals who might be considered *false positives* (e.g., persons who returned to work for their former employers, presumably doing the same work and not subsequently displaced again).
SDB17	During the period January 1995 through December 1997, did you leave a job or lose a job for any reason?
	Rationale: SDB17 was asked of all persons for whom a "no" answer was provided to supplement item SD1. The goal was to identify persons who might have been missed as displaced workers (see SDB20).]
SDB20	What is the MAIN reason you are no longer working at that job? [Note: This item had twenty-two response precodes, seven *employer-related reasons*) and fifteen *personal reasons* (see SDB3 for examples).
	Rationale: Generally speaking, to determine if the person lost or left a job involuntarily (i.e., one of the employer-related reasons) or voluntarily (i.e., one of the personal reasons). With respect to employer-related reasons only, this item was useful for identifying potential *false negatives*.

Table 11A. Debriefing Item SDB1

% (N=1342)	Earlier you told me that me that (name/you) had lost or left a job in the past three calendar years [see CK2-SDB1 for fill instructions]. Did (you/name) lose that job or did (you/she/he) leave that job?
62.2% (835)	<1> Lost job
35.1% (471)	<2> Left job
2.0% (27)	<3> Retired from job
0.3% (4)	<D> Don't know
0.4% (5)	<R> Refused
Objective:	For persons reported to have lost or left a job, to determine if the specified person lost their job or left their job.

Table 11B. Debriefing Item SDB3Z

% (N=897)	Did you ever return to work for that employer, for even a short period of time?
9.0% (81)	<1> Yes
90.9% (815)	<2> No
0.1% (1)	<D> Don't know
---	<R> Refused
Objective:	For persons reported to have **lost, left, or retired from** a job during the reference period for a displacement reason: To determine if the person returned to work for that employer, even briefly. This question is an attempt to identify individuals who might be considered "false positives" (e.g., persons who returned to work for their employers, presumably doing the same work and not subsequently displaced again). The only conclusion one might reasonably draw from a high percentage of 'yes' responses to this question (say 20% or more) is that supplement items SD1 and SD2 are not being interpreted as intended.

Table 11C. Debriefing Item SDB17

% (N=18372)	During the period January 1995 through December 1997, did (name/you) leave a job or lose a job for any reason?
13.4% (2458)	<1> Yes
85.6% (15729)	<2> No
0.5% (83)	<D> Don't know
0.6% (102)	<R> Refused
Objective:	For persons reported NOT to have lost or left a job during the reference period, to initiate a line of questioning that may be useful in determining if some of these persons actually did lose or leave a job for displacement-related reasons. Such persons are referred to as "false negatives".

Table 11D. Debriefing Item SDB20

% (N=2103)	What is the MAIN reason (you are) (she/he is) no longer working at that job? [CHECK **ONE** OPTION ONLY]
	Note to Interviewers: If the respondent provides multiple reasons for why the person is no longer working at that job, tell the respondent that we are looking for the MAIN reason she/he is no longer working at that job.
	Employer-Related Reasons
1.0% (20)	<1> Employer closed down business (or was about to close down business)
0.6% (13)	<2> Employer moved away (or was about to move away)
2.3% (48)	<3> Employer was downsizing or restructuring
1.7% (36)	<4> Employer had insufficient work
0.6% (12)	<5> Worker's position/shift was abolished (or was about to be abolished)
0.5% (10)	<6> Seasonal job completed
0.3% (6)	<7> Self-operated business failed
	Personal Reasons
5.9% (124)	<8> did not like job *or* boss
22.7% (477)	<9> better job / different job
7.7% (161)	<10> not enough PAY / to get more pay
0.6% (13)	<11> poor benefits / no benefits
6.6% (139)	<12> OWN illness/injury
4.9% (103)	<13> child care problems / family obligations
3.5% (73)	<14> maternity / pregnancy
1.0% (21)	<15> RETIRED
0.3% (7)	<16> left military service (e.g., Army, Navy)
1.8% (38)	<17> fired
10.8% (228)	<18> moved away
6.8% (143)	<19> school / training
2.8% (58)	<20> to start own business
1.1% (23)	<21> too long of a commute
15.6% (329)	<22> OTHER [**Specify:**_____]
0.7% (15)	<D> Don't know
0.3% (6)	<R> Refused
Objective:	To determine why the specified person lost or left a job. The first set of reasons refers to situations that would result in a displacement classification. With one possible exception (i.e., retired), the second set of options—including the 'other' option—would result in a non-displacement classification. However, it is possible that some personal reasons actually mask an employer's earlier attempt to let the worker go, so some personal reasons are channeled toward SDB22.

A summary of the methods and findings for the set of three DWS field tests can be found in **Table 12**. In retrospect, the first field test (1996) can best be described as exploratory. The second field test (1998) was far more comprehensive in that the scope of work was expanded to address issues that surfaced during the 1996 research and that had been raised during "forums" with subject-matter experts (i.e., labor force economists). The third field test (2000) replicated some of the work conducted during the prior tests but also sought to go beyond earlier work to evaluate a set of questions that might be useful in a *redesigned* supplement questionnaire.

Subsection 5.3. *Analyzing and Integrating Field-Test Evaluation Data in the Context of Survey Metadata and the Broader Questionnaire-Design-and-Evaluation Process*. Some individuals may believe that conducting a field test is about finding—and ultimately repairing—design problems with survey questions and the questionnaires in which they are embedded. At best, that belief would only seem valid in rare cases—specifically, when all of the developmental, design and evaluation work conducted prior to field test (e.g., at P6) has been done flawlessly. When conducting a field test, a survey practitioner is not only evaluating the survey questions that comprise a given questionnaire, but (in principal) also all of the documented-and-available observation-based knowledge/information/data and all of the available-and-documented prior evaluation work that informed the design of those questions (see Table 1). *Changes* occurring in the domain-of-interest and in the world-at-large (social, cultural and technological) only complicate an already complex process. Of course, some survey questions (and the questionnaires in which they are embedded) have obvious design flaws, easily corrected in some cases; but we delude ourselves, our sponsors and other interested parties if we assume that the question wording and sequencing are the *only* problems that require fixing when the data generated by a particular questionnaire does not appear to be "behaving" well.

Table 12: Summary of Methods and Findings for the Three DWS Field Tests (1996-2000)

Field Test	Comments (C), Methodological Details (D) and Illustrative Findings (F)
1996	▪ **C:** This phase can best be described as exploratory field test. This initial evaluation focused on two supplement items, SD1 and SD2.
Interviewer Debriefing	▪ **D:** One focus group involving 10 telephone center interviewers. ▪ **F:** Evidence of conceptual problems (e.g., what constitutes a job), cognitive problems (e.g., meaning of the phrase "or another similar reason"; difficulty with the distinction between losing and leaving a job) and design/operational problems (e.g., failure to read all parts of questions).
Interaction Coding	▪ **D:** 52 person interviews coded (behavior coding). ▪ **F:** Evidence of problems with interviewers reading SD1 and SD2 as worded (12% and 57% of cases with major changes, respectively); respondents also had difficulty providing adequate answers to SD2 (33% of cases had inadequate answers).
Respondent Debriefing	▪ **D:** Debriefing questionnaire consisting of 8 response-dependent probe questions. ▪ **F:** Evidence of possible displaced-worker undercount in the order of 25 percent (false negatives). About one-third of the suspected undercount was traceable to SD1, precode 6, and the remainder to inaccurate "no" answers to SD1 (unexplained).
1998	▪ **C:** Relative to the quality assessment work conducted in 1996, this second phase was far more comprehensive. Again, the evaluation focused on SD1 and SD2.
Interviewer Debriefing	▪ **D:** Three focus groups involving 34 telephone center interviewers. Interviewers were also asked to rate SD1 and SD2 in terms of how difficult they thought these items were for respondents to answer. ▪ **F:** Evidence of conceptual problems (e.g., what to do about temporary jobs and other alternative work arrangements), cognitive problems (e.g., uncertainty regarding the meaning of terms such as "insufficient work" and "layoff") and design/operational problems (e.g., awkward transition phrase in SD2; parents reporting for older children; burden on the elderly and the disabled; interruptions). Rating scale data (means and standard deviations) for SD1 and SD2 provided evidence of considerable variability within and between groups of telephone center interviewers.
Interaction Coding	▪ **D:** 145 person interviews coded (behavior coding). ▪ **F:** Evidence of problems reading SD1 and SD2 as worded (13% and 72% of cases with major changes, respectively); respondents also had difficulty providing adequate answers to both items (10% and 28% of cases had inadequate answers, respectively).
Respondent Debriefing	▪ **D:** Debriefing questionnaire consisting of 22 response-dependent probe questions. ▪ **F:** Evidence of possible displaced-worker undercount in the order of approximately 20 percent (false negatives). Again, about one-third of the suspected undercount was traceable to SD1, precode 6, and the remainder attributable to inaccurate "no" answers to SD1 (unexplained). However other debriefing data raises questions as to the actual status of some "displaced workers" (e.g., 23% of cases categorized as displacements due to "insufficient work" were later reported to have been temporary jobs); some labor force economists would exclude persons whose jobs were temporary from the count of displaced workers (potential false positives).

[Table 6 continues on the next page.]

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

We support some of these claims by reviewing findings presented earlier with regard to supplement items SD1 and SD2, the two supplement items that carry the bulk of the load for classifying target persons a displaced or not displaced from a job. Before doing so, and in fairness to the content specialists (i.e., labor force economists) who designed the original questionnaire in the early 1980s, it is important to note that the DWS not designed as a panel survey; its first administration in 1984 was supposed to its last. But, as it sometimes happens when a particular survey attracts a constituency of data users, the DWS has not only survived, it has been administered (with some design changes) biennially since it first appeared in 1984. In reviewing relevant metadata (Table 5) and findings for the 1998 field test (see Table 12 for a

very brief summary of evaluation data for all three field tests), let's follow the sequence in which the field-test methods were implemented: behavior coding (see Table 9), interviewer debriefing (see Tables 6, 7 and 8) and respondent debriefing (see Tables 10 and 11). The behavior-coding data reveal some serious issues with SD1 and SD2, and particularly the latter. For all three field tests, there were high percentages of interviewer problem codes (i.e., major wording changes) and of respondent problem codes (i.e., inadequate answers, interruptions), again, particularly for SD2. Monitoring and coding interviewer-respondent exchanges was particularly good preparation for conducting the three focus-group debriefings with CPS/DWS interviewers in 1998. Data in Table 7 suggest that some respondents had expressed uncertainty regarding interpretation of the phrase "some other reason". Interviewers, too, seemed to be uncertain as to what this phrase meant. A review of supplement metadata reveals a potential source of this uncertainty: The term "similar reasons" is defined (in part) as including "… 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 … (italics added)." To most laypersons (e.g., interviewers, respondents and myself), it is not entirely clear what that definition/description might mean or entail. Most interviewers apparently assumed that the phrase meant for a reason *similar to* one of the reasons specifically mentioned in the body of the question (SD1), which are all valid displacement reasons. Apparently that was not what the survey sponsors had intended (in 1998 anyway), because responses coded as "some other reason" in SD2 are skipped out of the DWS and those respondents are *not counted* as displaced workers. Debriefing data collected from respondents whose answer to SD2 was coded as "some other reason" (i.e., verbatim information regarding the specific reason for job loss; see **Appendix** Table A-4) would later suggest that about one-third of those respondents apparently did lose their

jobs for a displacement reason; these cases represent potential *false negatives*. The problems identified by interviewers with regard to SD2, such as the widespread failure to read the "read if necessary" statement (for wording, see Table 4)—How would interviewers even know to read this statement unless a respondent specifically mentioned at SD1 that she/he had lost more than one job during the past three years?—only compounded the problems with accurate classification. The ratings data gathered from interviewers during the course of the focus groups (Table 7) document the elevated level of difficulty some respondents appeared to be having with SD1 and SD2, and also differences in perceived difficulty levels among interviewers who work in different centralized CATI settings. Lastly, respondent debriefing data were useful with regard to exploring conceptual/substantive issues not fully addressed in the DWS metadata (Table 5) and with regard to computing a crude estimate of the level of measurement error associated with these two supplement items. For example, with regard to the former, some labor force economists wanted to know what percentage of displaced workers had "left a job" as opposed to having "lost a job" (see **Table 11A**), believing perhaps that not all job leavers should not be counted as displaced workers. Regarding the latter point above, other debriefing questions took direct aim at measurement error, with some focusing on false positives (see **Table 11B**) and some focusing on false negatives (see **Tables 11C and 11D**). For example, with regard to the latter, persons who had answered "no" to SD1 (i.e., persons who said they did not lose or leave a job during the reference period for a displacement reason) were later asked if they had lost or left a job *for any reason* during that time (see Table 11C)—about 13 per cent had done so. Those persons were then asked why they were no longer working at that job, and about 7 per cent gave a response that could be coded as a displacement reason. Though these percentages may seem small, the number of *potential false* negatives is actually quite substantial

when multiplied by the total number of "no" response to SD1. And when one considers other debriefing data (i.e., verbatim data gathered from respondents whose answers to SD2 were coded as "some other reason" (see **Appendix**, Table A-4), the estimate of false negatives (i.e., persons not counted as displaced workers who probably should have been) approaches 20 percent. That said, we must add an important caveat: The actual level of measurement error depends greatly on how one interprets the relatively lean (and sometimes vague) metadata for the displaced worker supplement (see Table 5) and how motivated respondents might have been in providing accurate/honest response to the debriefing questions.

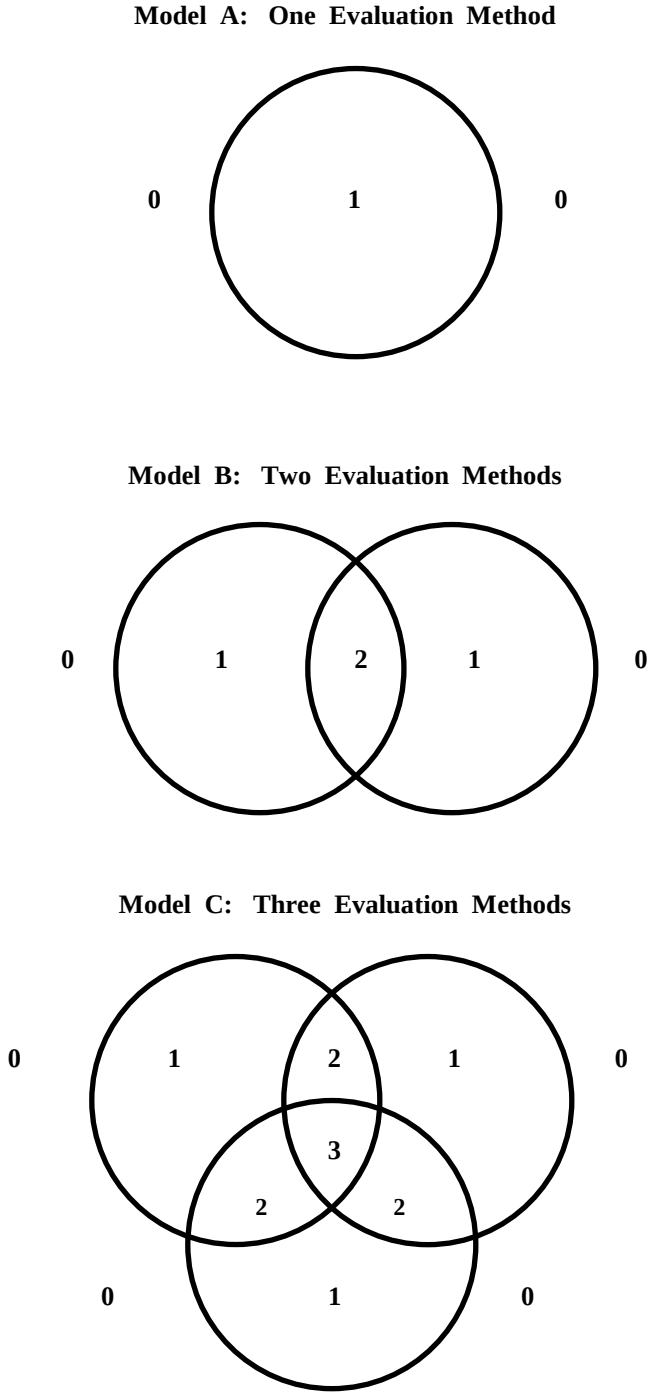
In the case of the DWS, integrating findings from the various evaluation methods/techniques and forming a judgment as to how serious the measurement-error issues might be for SD1 and SD2—the principal questions used for classification purposes—was not particularly difficult; but that is not always going to be the case. When using multiple methods to evaluate a questionnaire, there are always going to be situations in which data from one method appears to be inconsistent (e.g., no apparent problems) with the data from others (e.g., lots of problem indicators). Moreover, one might have a "bias" towards some methods (e.g., respondent debriefing using follow-up probe questions) relative to others (e.g., behavior coding; interviewer debriefings using a focus-group format). There are no magic formulas for integrating research findings from various evaluation methods, to my knowledge. Some practitioners favor quantitative data (e.g., generated from follow-up probe questions) over qualitative data (e.g., focus group commentary), and in some cases there may be good reasons for doing so; however, in other cases, there may be good reasons for not doing so (e.g., when follow-up probe questions are not particularly well designed or when they are not designed to gather data on *both* potential false negatives and potential false positives). When there are no compelling reasons to favor

data from one method over data from other methods in a multiple-method evaluation effort, one looks for "consensus" (or the lack thereof) among the evaluation methods. We tend to rely on a common-sense approach that we have entitled *the relative confidence model* (Esposito and Rothgeb, 1997, pp. 563-565).

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

As present (2009), nine years after the last of three DWS field tests (2000), various issues regarding the conceptualization and the measurement of job displacement remain.

Figure 1. A Relative Confidence Model for Identifying Problematic Survey Questions



Subsection 5.4. *A Subjective Assessment of the Utility of the Various Evaluation Methods and Techniques Used in the DWS Field Tests (**Table 13**).* The three principal evaluation methods used in the present research effort attempt to capture or reveal the perspectives of various informational sources (see column headings of Table 1, *Interdependent Sources of Measurement Error*). 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., confusing or inadequate item specifications; poor question design; inappropriate or insufficient 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.

Table 13. A Subjective Assessment of the Utility of Various Questionnaire Evaluation Methods and Techniques Used in the DWS Field Tests

Methods	Comments and Observations
Interviewer Debriefing	
▪ Log Forms	▪ Useful in preparing focus-group participants for documenting and reporting problems experienced with various questionnaire items. ▪ Not all interviewers seem motivated to contribute/participate fully. ▪ Whenever there is a real or a perceived conflict between research tasks and production goals/objectives, the latter take priority at the expense of the former.
▪ Focus Groups	▪ Qualitative data: Retrospective and subject to situational effects (e.g., group dynamics). ▪ Useful for identifying conceptual and operational problems. ▪ CATI interviewers not necessarily representative of population. ▪ Provides no quantitative basis for estimating measurement error. ▪ Useful in corroborating or contradicting behavior-coding observations.
▪ Rating Form/Scale	▪ Descriptive quantitative ratings data: Retrospective and potentially contaminated if interviewers talk about items prior to completing the rating task. ▪ Useful in identifying differences among interviewers, but sample of interviewers not representative of population. ▪ Minimal labor on part of researcher. ▪ Provides no quantitative basis for estimating measurement error.
Interaction Coding	
▪ Behavior Coding (live, low-tech coding)	▪ Descriptive quantitative data and some qualitative data. ▪ Useful in detecting possible problems with specific items, but not necessarily useful in identifying solutions. ▪ Useful for comparative analyses (open vs. closed questions) ▪ Unobtrusive. ▪ Relatively objective/unbiased. ▪ Sample of interviewers and respondents not fully representative of their respective populations. ▪ Live coding more susceptible to error and omissions than other coding strategies (e.g., coding from audiotapes) ▪ Provides no quantitative basis for estimating measurement error.
Respondent Debriefing	
▪ Response-dependent Follow-up Probes	▪ Quantitative data: Useful in confirming/quantifying specification problems (see last bullet in this set). Data rich in that, as need arises, cross-tabulations can be run with other debriefing items and with items from the host questionnaire. ▪ Qualitative data: "Other-specify" precodes provide quasi-ethnographic data. ▪ Respondent sample presumed to be fairly representative of population, but this not necessarily the case. ▪ Probes can increase respondent burden in some cases. ▪ Labor intensive for content and design specialists. ▪ Potentially very useful in estimating measurement error associated with specific items. However, potentially misleading if questions are not balanced with respect to identifying false positives and false negatives.

As other practitioners have noted, each of these techniques possesses certain weaknesses. With regard to the use of follow-up probes, it is not always clear what probe questions one might need to ask and, even when an objective for a probe is clear, one may not be completely successful in achieving that aim. For example, in the 1998 field test, a debriefing question was asked to determine if the job a person lost or left for a displacement reason was a temporary job: "Was the job you lost a temporary job, that is, a job that was supposed to last only for a limited time or until the completion of a project?" The expectation was that a large majority of persons for whom a "yes" answer was provided would have worked at such jobs for relatively brief periods of time (e.g., six months or less). When the debriefing item was cross-tabulated with a supplement item on employment duration (n=108), it was found that approximately 40 percent of displaced workers had worked for their employer for more than a year and that 25 percent had worked for more than two years. In other words, probe questions can be just as problematic as the questionnaire items they are designed to evaluate. With regard to interviewing debriefing techniques, focus groups are highly susceptible to group dynamics and, depending on how research participants are selected, may not be representative of the interviewer population. Retrospective rating forms are subject to memory or salience effects, and occasionally yield findings that are difficult to explain. For example, in phase two, interviewers at two telephone centers had identified numerous problems with SD2; when asked to rate this item, 12 of 22 interviewers gave it relatively high difficulty ratings (3-to-5 range). Quite inexplicably, not one interviewer in a group of twelve at the third telephone center identified SD2 as problematic (see Table 7) and, as a result, SD2 was not rated at that location. With regard to behavior coding, the principal weakness associated with this technique—given the manner in which we chose to

employ it—is that, while it is useful in identifying where problems exist, it provides little guidance as to what may be causing these problems. Another weakness—associated more with the coder (the present author) than with the technique—was that only interviews with English-speaking respondents could be monitored and coded.

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

SECTION 6. Closing Remarks

There is relatively little to prevent survey organizations from doing suboptimal questionnaire design-and-evaluation work, *other than* the expertise and professionalism of the men and women who populate the survey organizations and do the work. And because of the substantial resources invested in their implementation and of the expectations regarding the importance of their findings, field tests are especially critical.

Subsection 6.1. *The Collaborative Nature of Field Tests from a Survey Practitioner's Perspective.* For field tests to be successful, expertise, communication and collaboration are essential. Content specialists need to know their subject-matter domains and communicate that knowledge/information/data to others in a clear and understandable manner. Design specialists need to be familiar with the domain-of interest and relevant metadata, understand questionnaire-design principles and be proficient using the methods available for evaluating questionnaires. Interviewers need to be properly trained and adequately compensated for the challenging work they perform. And respondents need to be respected and encouraged to provide accurate information/data about behaviors that have implications for sustaining and enhancing the common good. To ensure that field tests provide useful, high-quality data, all parties must understand their roles and perform them faithfully and competently. Such undertakings will not necessarily fail if a relatively small number of interviewers and respondents choose to satisfice or to disengage, but is will not survive for long as a viable collaborative process if content specialists and design specialists are not knowledgeable, competent and fully engaged.

Because of what we assume, believe and think we know about the various phases of questionnaire design-and-evaluation process and the interdependencies among its various collaborators, survey practitioners have a special responsibility to monitor the functioning of the

entire design-and-evaluation process. If content specialists are vague when specifying concepts or question objectives, we need to request that they be more precise; if they appear to be biased or uninformed about various aspects of their subject-matter area, we need to encourage them to expand their knowledge base. If interviewers are inexperienced or uncertain as to their proper roles, we need to insist that they receive the proper training. If respondents seemed overwhelmed with the sheer number of questions we ask, or seem confused by the content or structure of the questions we have drafted, we need to take steps to minimize burden, clarify question content and, as necessary, take steps to upgrade our question/questionnaire design skills. If there is persuasive evidence that the meaning conveyed by the questions we have designed does not match the sponsor's intent or what appears to be the case in the world-at-large, we need to make sure that such situations are rectified. Such tasks circumscribe our role as practitioners (and as professionals)—and, as such, represent the work we need to do and should be expected to do.

Subsection 6.2. *Incorporating Field-Test Research Findings within Q-Bank*. Any "database management system" that can be used to capture, characterize and organize the questionnaire-evaluation research being conducted by survey practitioners within the federal government (and elsewhere) would be a welcome and a very significant contribution to our profession, in my view. The teams of survey practitioners responsible for the current status of the Q-Bank system are to be commended for their efforts and their accomplishments thus far—the computer-based system is quite impressive in its scope and detail (National Center of Health Statistics, 2009). Having now reviewed available documentation for Q-Bank (e.g., user's manual for interviewer-administered population questionnaires) and read various papers that have been written on the system (e.g., Beatty, Willis, Hunter and Miller, 2005; Bradburn; 2005), it does appear that the

comprehensive coding system—developed originally for reporting findings from cognitive interviews—is flexible enough to incorporate findings from multiple-method field tests such as the two efforts discussed in the current paper. One needs to recognize however, that what practitioners have available to them in terms of metadata (e.g., conceptual specifications; interviewing manuals; prior research findings) is likely to be substantially greater for field tests (e.g., at P6 or P8) than for cognitive interviews (e.g., at P4); and this fact has implications both for those who contribute to the system (and who are expected to provide links to relevant conceptual documentation and prior research findings) and those who make use of the system (and who, in theory, would want and need to read that metadata in order to understand why there are "issues" with a given questionnaire item). To accommodate field tests, some of the descriptions of the Q-Bank database fields would have to be expanded: For example, with regard to *interviewer instructions* (see p.19 of the user's manual for interviewer administered questionnaires), as a means of assessing the utility and clarity of question-specific instructions that have *or have not* been made available in survey-specific interviewer training materials (e.g., memoranda or manuals). Other database fields may need to be added (or an existing one modified) to accommodate what can be accomplished with certain methods; for example, well-designed post-interviewer follow-up probes used to debrief respondents during P6 or P8 field tests may provide crude but potentially revealing quantitative evidence for the existence of *measurement error* in a survey's key estimates—and it is not apparent how such data are to be coded. When incorporating any of the various methods that practitioners might employ in conducting a field test, one must also be concerned about the *variants* of these methods. For example, in their comprehensive review of the behavior-coding literature, Ongena and Dijkstra (2006) identified 48 different schemes for coding interviewer and respondent interactions during

questionnaire administration. In so doing, they identified which behavior codes were used most commonly (e.g., on the *interviewer* side: question read exactly as worded, question read with major change; and on the *respondent* side: adequate answer, inadequate answer; refusal to answer), but clearly not universally. How then should behavior-coding data be incorporated within Q-Bank: Should all practitioners who have used this method report their findings in terms of these most common codes or should they simply describe their coding system and provide data for those codes? If the former approach is adopted, who determines the rules for converting the more complex coding systems to the simpler, common-denominator coding system? If the latter, does the lack of comparability across research efforts become an issue? What is true of behavior coding in terms of standardization of application is true of other methods, like interviewer debriefing (e.g., using focus groups) and respondent debriefing (e.g., using follow-up probes) as well; and what does and does not get reported as data is not always apparent or transparent. The more methods employed in any one field test, the more challenging the system becomes for Q-Bank developers, contributors and users alike, and the more compelling Norman Bradburn's (2005) sage counsel regarding successful questionnaire design-and-evaluation database systems: *simplicity* in development and use.

APPENDIX

Table A-1. CPS Redesign (First Field Test, 1990-1991): Research Leading Up to the Selecting of the "Work" Question for the New CPS Questionnaire

Alternative Questions

- Version A: Did you do any work at all LAST WEEK, not counting work around the house?
- Version B: LAST WEEK, did you do any work for pay or profit?
- Version C: LAST WEEK, did you do any work at all? Include work for pay or other types of compensation?

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

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

Methodological Findings

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

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

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

- *Focus groups*. Some interviewers report not liking the A question because it sounds demeaning to housewives and because it is confusing to some respondents (e.g., volunteer workers). The use of the term "profit" in the B question confuses some respondents--especially those who do not have a business. The use of the phrase "other types of compensation" in the C question confuses some respondents and some interviewers, too.
- *Debriefing questionnaire* (*N*=68 interviewers). When asked what question was most difficult for them to ask, two interviewers selected the version A work question (too wordy or awkwardly worded), three selected the version B work question (confusing, ambiguous, difficult to understand), and three selected the version C question (same reasons as B). When asked what question appeared to be most difficult for respondents to answer, five interviewers selected the version B work question (confusing, ambiguous, difficult to understand) and three selected the version C work question (same reasons as B). When asked what terms or concepts were most commonly misunderstood by respondents, six interviewers mentioned "working for pay or other types of compensation"; four mentioned "working for pay or profit" or just "profit"; and four mentioned "work" or "work vs. employed".

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

- *Follow-up probe questions.* All three work questions were effective at identifying employed persons. Differences in the percentage of employed *individuals missed* for all possible question pairings were not significant (A=2.0%; B=1.8%; C=1.1%).
- *Vignettes.* No one of the three work questions was better at eliciting responses that match CPS definitions (i.e., no one question clearly outperformed the other two alternatives). Some evidence to suggest that version B question wording may be less inclusive than other alternatives, in that higher percentage of respondents say "no" to all vignette scenarios. Version B question is less successful than alternatives in correctly classifying marginal work activities (e.g., work in the home), but better at correctly classifying non-work activities (e.g., volunteer service).

[Table A-1 continues on the next page.]

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

- *Response-distribution analyses.* All three work questions produced approximately the same percentage of individuals reported as working (A=59.16%; B=57.95%; C=58.71%; differences in stated percentages for all possible question pairings are not significant).
- *Nonresponse analyses.* Very little item nonresponse across versions (A=0.18%; B=0.18%; C=0.22%).

Recommendation and Justification

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

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

Table A-2. CPS Redesign: Selected Results for the "Work" Question Across Three Fields **

Test/ Design	Dates	Sample Size	Q'aire Version	Interviewer Debriefing	Respondent Debriefing	Behavior Coding		Response Distribution
					% Missed Employment	*INT Code (% E + mC)*	*RSP Code (% AA + qA)*	*% Yes (% NR)*
Field Test 1 CATI/RDD	July 1990-January 1991	70,000 HHs Cumulative All Versions	A	see Table A-1	5.5% (5.0% paid)	94% CATI	91% CATI	59.16% (0.18%)
			B	see Table A-1	2.3% (1.5% paid)	99% CATI	96% CATI	57.95% (0.18%)
			C	see Table A-1	1.6% (1.0% paid)	94% CATI	92% CATI	58.71% (0.22%)
Field Test 2 CATI/RDD	July 1991-October 1991	32,000 HHs Cumulative Both Versions	A	---	3.8% (2.2% paid)	---	---	57.74% (0.15%)
			D	FG: "just my job"	2.6% (2.0% paid)	100% CATI	95% CATI	57.01% (0.08%)
Field Test 3 CATI/CAPI Address List Sample	July 1992 to December 1993	144,000 HHs Cumulative (1993 only)	New CPS Question-naire	FG: 35.5% IDQ: 18.2%	2.9% (1.6% paid)	100% CATI 99% CAPI	98% CATI 93% CAPI	58.58% (0.16%)

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

Table A-3. Log Form for Keeping Track of Problems with Key Supplement Questions

KEY QUESTIONS and LOG SHEET
for Telephone Center Interviewers Participating in Focus Groups

INSTRUCTIONS FOR TELEPHONE-CENTER INTERVIEWERS who will be serving as focus group participants in February 1998: On the attached LOG SHEET, please keep a record of any problems that you (or respondents) may be experiencing during the administration of the supplement, especially with regard to the key questions that appear below. **Please *bring these sheets* with you when you come to the focus group session.**

Item Label	Key Supplement Items
SD1	During the last 3 calendar years, that is, January 1995 through December 1997, 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?
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. <1> Plant or company closed down or moved Plant or company operating but lost or left job because of: <2> Insufficient work <3> Position or shift abolished <3> Seasonal job completed <5> Self-operated business failed <6> Some other reason
SD3	In what year did you last work at that job?
SD4	(Do/Does) (you/he/she) expect to be recalled to that job within the next 6 months?
SD5	Had (name/you) been given written advance notice informing (you/him/her) that (the plant or business would be close) ((you/he/she) would lose (your/his/her) job)?
SD6	How long before (you/he/she) (were/was) to have lost (your/his/her) job did (you/he/she) receive that notice?
SD7	(Were/was) (you/name) employed by government, by a private company, a non-profit organization, or (was/were) (you/he/she) self-employed or working in a family business?
SD18	How long had you worked for (fill job) when that job ended?
SD25	After that job ended, how many weeks went by before you started working again at another job?

[Table A-3 continues on the next page.]

Table A-3. (continued)

[Log Form, page 2]

LOG SHEET for Focus Group Participants [Telephone Center Interviewers]

INSTRUCTIONS: Please use this log sheet to identify **supplement items** that are causing problems for you or respondents during the administration of the supplement. (Use back of sheet, or add sheets, if more space is needed.) Additional information on problem types and a sample log entry are provided below for illustrative purposes. Please include the ***item name*** (e.g., "ST1") when describing a particular problem.

The general types of problems interviewers might encounter include the following:

- ***question-specific problems,*** such as when interviewers have difficulty coding a respondent's answer to a particular item
- ***comprehension problems***, such as when the respondent has difficulty understanding a particular question or the specific words/terms used in that question
- ***proxy problems***, such as when proxy respondent appears to be guessing at answers to a particular question
- ***response problems***, such as when respondents refuse to answer a particular question and/or refuse to finish the supplement because of the sensitivity of the information being requested

Sample Log Entry
SD*#*—(*need full item name here*): The respondent seemed to be having a problem with (*fill with the word that appears to be causing problems*). She asked me twice to tell her what that word meant. Also, this question is way too l-o-o-o-n-g! She interrupted with an answer before I could finish reading the question.

Table A-4. Verbatim Entries to Debriefing Items SDB3S and SDB20S

Listed below are verbatim entries from two respondent debriefing items that illustrate the types of responses that interviews may have difficulty coding into one of the precoded displacement categories (see supplement item SD2, options 1-3, in the current DW/JT questionnaire) or one of the non-displacement categories (options 4-5). SDB3S lists the types of entries interviewers code as option 6 ("other") in SD2 of the current DW/JT supplement. SDB20S lists the types of entries that we might classify as *false negatives* (i.e., persons for whom a "no" answer was provided to SD1, but who actually may have been displaced from a job).

List 1. Verbatim Entries from SDB3S: *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?*

1. Works in construction; when one job finishes, he moves to the next
2. No work, slack work
3. Lack of work
4. Company merged with another company
5. Lack of funding
6. Another bank bought out bank that person was working for
7. Laid off permanently
8. Employer cut person's hours
9. Employer sold business—new owner didn't need workers
10. Office closed and had to move
11. Because of the Asian stock market crash
12. Pushed out of position
13. Another company took over regional hospital
14. Bank was bought out so she lost her position
15. Program was not refunded
16. Dispute with management, taken over by new management
17. Company couldn't afford her services anymore
18. Business was sold

List 2. Verbatim Entries from SDB20S: *What is the MAIN reason you are no longer working at that job?*

1. New ownership
2. Company under new management
3. New owners took over the business and fired person
4. On strike for three weeks
5. Relocation—was hired three months later
6. Renegotiated contract did not include commissions—so person said "no"
7. Never called back to work
8. Heard rumors that position was being eliminated—and it was
9. Reduced wages and restructuring of waitress [duties (?)]
10. Company contracted for Department of Energy; funding was unstable
11. Company was part of acquisition by other company
12. Employer was trying to get rid of experienced staff
13. Person was out on workers compensation and employer wouldn't hire him back
14. Laid off
15. To keep job classification, would have to relocate

C:\DWJT\98\RD\Verbatims.doc 111798

References

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

Barsalou, L.W. (2008). What is Grounded Cognition? *Annual Review of Psychology*, 59, pp. 617-645.

Beebe, J. (2001). *Rapid Assessment Process: An Introduction*. Walnut Creek, CA: AltaMira Press.

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

Beatty, P., Willis, G., Hunter, J.E. and Miller, K.M. (2005). Design of the Q-Bank: Determining Concepts, Content, and Standards. *ASA Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 981-988

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

Belli, R.F., Lee, E.H., Stafford, F.P., and Chou, C.H. (2004). Calendar and Question-List Survey Methods: Association between Interviewer Behaviors and Data Quality. *Journal of Official Statistics*, 20, pp. 185-218.

Biemer, P. (2004). Modeling Measurement Error to Identify Flawed Questions. In S. Presser, J. Rothgeb, et al. (eds.) Methods *for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 225-246.

Bradburn, N. (2005). Comments on Q-Bank Papers. *ASA Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.

Campanelli, P.C., Martin, E.A., and Rothgeb, J.M. (1991). The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data. *The Statistician*, 40, 253-264.

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

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

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, pp. 475-495.

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

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

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

DeMaio, T.J., and Landreth, A. (2004). "Do Different Cognitive Interviewing Techniques Produce Different Results?" In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 89-108.

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.

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, pp. 909-918.

Esposito, J.L. (2005). Primum non Nocere: An Oath for Survey Practitioners? *QUEST2005: Proceedings of the Fifth Workshop on Questionnaire Evaluation Standards*. Heerlen: Statistics Netherlands, pp. 151-165.

Esposito, J.L. (2004a). With Regard to the Design of Major Statistical Surveys: Are We Waiting Too Long to Evaluate Substantive Questionnaire Content? In P. Prüfer, M. Rexroth, F.J. Fowler, Jr. (eds.), *QUEST 2003: Proceedings of the Fourth [Workshop] on Questionnaire Evaluation Standards*. Mannheim, Germany: Zentrum für Umfragen, Methoden und Analysen (ZUMA), pp. 161-171.

Esposito, J.L. (2004b). Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study. *Journal of Official Statistics*, 20, pp.143-183.

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

Esposito, J.L., and 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, pp. 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.

EUROSTAT (2006). *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System.* Unpublished document.

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.

Forsyth, B., Rothgeb, J.M., and Willis, G.B. (2004). "Does Pretesting Make a Difference? An Experimental Test." In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 525-546.

Fowler. F.J (1995). The Case for More Split-Sample Experiments in Developing Survey Instruments. In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 173-188.

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

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

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

Fowler, F.J. and Cannell, C.F. (1996). Using Behavior Coding to Identify Cognitive Problems with Survey Questions. In N. Schwarz and S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey-Bass, pp. 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, pp. 217-234.

Glaser, B.G., and Strauss, A.L. (1967/1999). *The Discovery of Grounded Theory*. New York: Aldine de Gruyter.

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

Grice, H.P. (1975). Logic and Conversation. In P. Cole and J.L. Morgan (eds.), *Syntax and Semantics.* New York: Academic Press, pp. 41-58.

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, pp. 389-402.

Groves, R.M. (2006, Special Issue). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70, pp. 646-675.

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.

Groves, R.M., and Peytcheva (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, 72, pp. 167-189.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: Wiley-Interscience.

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, pp. 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, pp. 47-69.

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

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

Lessler, J.T., and Forsyth, B.H. (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, pp. 259-291.

Lindström, H., Davidsson, G., *Henningsson, B.,* Björnram, A., Marklund, H.*, Denell, C., Hoff, S.* (2004/2001). *Design Your Questions Right: How to Develop, Test, Evaluate and Improve Questionnaires.* Örebro, Sweden: Statistics Sweden.

Martin, E. (2004). Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation. In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 149-171.

Martin, E., Hunter-Childs, J., DeMaio, T., Hill, J., Reiser, C., Gerber, E., Styles, K., and Dillman, D. (2007). *Guidelines for Designing Questionnaires in Different Modes*. Washington, DC: U.S. Census Bureau.

Miller, K. (2001). "Making the Sponsor—Respondent Link in Questionnaire Design." *QUEST2001: Proceedings of the Third [Workshop] on Questionnaire Evaluation Standards*. Washington, DC: U.S. Bureau of the Census, pp. 92-98.

Morgan, D. L. (1988). *Focus Groups as Qualitative Research*. Qualitative Research Methods, Series 16. Newbury Park, CA: Sage.

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

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

National Center of Health Statistics (2009). Q-Bank: Improving Surveys through Sharing Knowledge. *Homepage for Q-Bank Website*. Hyattsville, MD: NCHS. http://wwwn.cdc.gov/qbank/Home.aspx

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

Olsen, K. (2006, Special Issue). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, 70, pp. 737-758.

Ongena, Y.P., and Dijkstra, W. (2006). Methods of Behavior Coding of Survey Interviews. *Journal of Official Statistics*, 22, pp. 419-451.

Platek, R. (1985). Some Important Issues in Questionnaire Development. Journal of Official Statistics, 1, pp. 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, pp. 73-104.

Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (Eds.) (2004). *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience.

Reeve, B.B., and Mâsse, L.C. (2004). Item Response Theory Modeling for Questionnaire Evaluation. In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 225-246.

Rothgeb, J.M., Loomis, L.S., and Hess, J.C. (2001). "Challenges and Strategies in Gaining Acceptance of Research Results from Cognitive Questionnaire Testing." *QUEST2001: Proceedings of the Third [Workshop] on Questionnaire Evaluation Standards*. Washington, DC: U.S. Census Bureau, pp. 79-89.

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.

Royston, P.N. (1989). Using Intensive Interviews to Evaluate Questions. In F.J. Fowler (ed.), *Health Survey Research Methods*, DHHS Publication Number PHS 89-3447. Washington, DC: Government Printing Office, pp. 3-7.

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

Schaeffer, N.C. (1991). Conversation with a Purpose—or Conversation? Interaction in the Standardized Interview [Chapter 19]. 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, pp. 367-391.

Schaeffer, N.C. (2002). Conversation with a Purpose—or Conversation? Interaction in the Standardized Interview [Chapter 4]. In D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds.), *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. Hoboken, NJ: Wiley-Interscience, pp. 95-123.

Schaeffer, N.C. and Dykema, J.L. (2004). Improving the Clarity of Closely Related Concepts: Distinguishing Legal and Physical Custody of Children. In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 475-502.

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

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

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.

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

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

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

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

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

Thomas, R. (1997). "The Questionnaire Development Environment and Its Implications for the Improvement of Questionnaire Design." *QUEST1997: Proceedings of the First Workshop on Questionnaire Evaluation Standards*. Örebro, Sweden, pp. 72-77. [Originally called the "MIST" Workshop: Minimum Standards in Questionnaire Testing.]

Tourangeau, R. (2004). Experimental Design Considerations for Testing and Evaluating Questionnaires. In S. Presser, J. Rothgeb, et al. (eds.) Methods *for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 209-246.

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, pp. 73-100.

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

Tucker, C., Brick, J.M. and Meekins, B. (2007). Household Telephone Service and Usage Patterns in the United States in 2004: Implications for Telephone Samples. *Public Opinion Quarterly*, 71, pp. 3-22.

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

U.S. Bureau of the Census (1998). CPS Field and CATI Interviewer Memoranda for the Displaced Worker Supplement, Number 1998-02. 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.

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

Willimack, D.K., Lyberg, L., Martin, J., Japec, L., Whitridge, P. (2004). "Evolution and Adaptation of Questionnaire Development, Evaluation and Testing Methods for Establishment Surveys." In S. Presser, J. Rothgeb, et al. (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 89-108.

Willis, G.B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

Willis, G.B, and Lessler, J. (1999). The BRFSS—QAS: A Guide for Systematically Evaluating Survey Question Wording. Rockville, MD: Research Triangle Institute.

Willis, G.B., Royston, P., and Bercini. D (1991). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. *Applied Cognitive Psychology*, 5, 251-267.

Paper presented as part of the "Workshop on Questionnaire Evaluation Methods", 21-23 October 2009, at the National Center of Health Statistics, Hyattsville, MD. Final paper submitted June 2010.