

“Primum non nocere”: An Oath for Survey Practitioners?

James L. Esposito

Bureau of Labor Statistics, Washington DC, USA.

The translation of the Latin phrase that appears in the first part of the title is: “First, do no harm.” Laypersons who read or hear that sentence may recognize it as the central imperative of the Hippocratic Oath. Of course, if a person believed that to be so, she/he would be mistaken—this imperative does not appear in the oath attributed to Hippocrates. (Some scholars attribute the statement to the Roman physician, Galen). Analogously, some of us may believe that conducting presurvey or postsurvey evaluation work and designing or modifying survey questionnaires on the basis of that work necessarily improves data quality. At best, given current practices, I suspect that such a belief is only partially valid and I will briefly review some methodological research (mostly mine) in an effort to support that claim. Regardless of the true origins of the imperative (Hippocrates, Galen, someone else), it is one that survey practitioners (and sponsors) should consider adopting as a guiding principle in the design and evaluation of survey questionnaires.

1. Introduction

Since its origins in 1997, the QUEST community has been responsible for some noteworthy contributions to the literature on questionnaire-evaluation standards and also appears to have had a significant impact on the professional practice of its members. The QDET conference brought together 338 attendees, incorporated 76 papers, and ultimately spawned a 25-chapter Wiley monograph (Presser, Rothgeb, Couper, Lessler, Martin, Martin and Singer, 2004, pp. xiv-xv) and a special issue of the *Journal of Official Statistics*. Various members of the community have independently or collaboratively published *books* (e.g., Presser, Rothgeb et al., 2004; Willis, 2005), *journal articles* (e.g., Akkerboom and Dehue, 1997; Haraldsen, 2004; Potaka and Cochrane, 2004), *book chapters* (e.g., Beatty, 2004; DeMaio and Landreth, 2004; Forsyth, Rothgeb and Willis, 2004; Fowler, 2004; Willimack, Lyberg et al., 2004), *best-practices/methodological monographs* (e.g., DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Lindström, Davidsson, Henningsson, et al., 2001/2004; Prüfer, Rexroth and Fowler, Jr., 2004), *conference/workshop papers* (e.g., Beukenhorst, Giesen, and de Vree, 2001; Cosenza and Fowler, 2001; Gower and Haarsma, 1997; Miller, 2001; Prüfer and Rexroth, 1999; Rothgeb, Loomis and Hess, 2001) and *other scholarly works* (e.g., Snijkers, 2002) on both household and establishment surveys. The QUEST workshops have provided a unique forum for some incredibly stimulating ideas in this very specialized research area. The whole experience has been exhilarating to some of us, if not professionally addictive. As a body of practitioners, we have learned a great deal. Yet, one thing has become apparent, painfully so in some cases: We still have much to learn—and not just with respect to the more-technical aspects of our craft (Thomas, 1997). Oftentimes, it seems, we are asked to make contributions to the design or evaluation of a particular survey only to find that the undertaking is grossly underfunded or the timeline for completing the work is impossible, or both. Some of us are left with little choice but to participate in such undertakings knowing full well that our design-and-evaluation work will be viewed as incomplete/ambiguous or, in a worst case scenario, as inaccurate or seriously flawed. As resources available for evaluation research dwindle, we can expect to be placed in these sorts of uncomfortable/untenable situations with increasing frequency. This paper describes a case study of one such situation in the hope that it will stimulate discussion on how members of the QUEST community might

effectively deal with such situations. As a general guideline, I will suggest the following: “Primum, non nocere.”

2. Additional Background Information

My heightened sensitivity to these low-resource-type research projects should be viewed in the context of prior experience with long-term, multiple-phase, design-and-evaluation research that for the most part has been well-supported, well-funded and well-staffed (Esposito and Rothgeb, 1997; Esposito, 2004a). This prior work dealt with important labor force issues (e.g., employment and unemployment; worker displacement) and carried significant policy implications. The case study to be described below can best be described as opportunistic; its sponsors did not possess the time or the funding for an elaborate design-and-evaluation effort. To their credit, they made the most of the limited resources they could muster. That said, I must confess to a not-so-latent socio-perceptual bias regarding sources of measurement error: Recent research has made me acutely sensitive to disparities in power among the various actors/participants who collectively represent the survey-data-collection enterprise (i.e., sponsors, subject-matter specialists, design-and-evaluation specialists, production specialists, interviewers and respondents). When problems arise with respect to data quality, too often it seems, the blame-attribution process seems to point in the direction of those participants who possess the least power—interviewers and/or respondents. In sociology and social psychology, this phenomenon is known as “blaming the victim.” This is not to say that interviewers and respondents should be viewed as innocent victims. They are not innocent, usually—they do misbehave, some more than others. However, this bias of mine compels me to focus more on *other explanations* as to why survey data quality is not as good as it can be. If successful, this paper (and the case study described below), will help to identify some of these “other explanations” (i.e., other sources/causes of measurement error).

3. A Case Study: The Cell-Phone-Use Supplement

3.1. Rationale and Objectives

This case study relates to the development and evaluation of a supplemental survey to the Current Population Survey (CPS), one of two primary labor force surveys conducted monthly in the United States. The 2004 cell-phone-use supplement was sponsored jointly by the Bureau of Labor Statistics (BLS) and the Bureau of the Census (BOC). The *rationale* for developing the supplement was a growing concern about the validity of certain types of telephone surveys (e.g., RDD surveys). One cause for concern was a lack of knowledge about that part of the population that national statistical surveys were not reaching—persons living in cell-phone-only households—and how the characteristics of persons in those households differ from the characteristics of persons in other households. A second cause for concern was that statistical agencies and private survey organizations are having more and more trouble reaching landline-telephone households. This supplement was designed to provide information on patterns of telephone usage in these households, especially how those households with both landline telephones and cell phones use the two technologies.

The primary *statistical objective* of the cell-phone-use supplement is to obtain estimates of four basic categories of telephone service available to and presently consumed by American households: (a) landline telephone service only; (b) cellular phone service only; (c) both landline telephone service and cellular phone service; and (d) no telephone service.

3.2. Supplement Questionnaire Development

The first draft of the supplement questionnaire was developed by a group of subject-matter experts (telephone survey methodology) from government, academia, and the private sector using items drawn from existing surveys conducted independently by researchers at Georgia State University and Arbitron (Tucker, Brick, Meekins and Morganstein, 2004). It is not known (to the present author) whether these borrowed items were accompanied by item-specific *metadata* (e.g., definitions of key concepts; item objectives). Later drafts of the questionnaire were refined on the basis of several rounds of cognitive testing conducted by private-sector researchers.

3.3. Evaluation Research

The plan for evaluating the cell-phone-use supplement involved both presurvey and postsurvey evaluations (pretesting and quality assessment, respectively).^{*} As noted, the draft supplement questionnaire was subjected to three rounds of cognitive testing (i.e., cognitive interviews with embedded topical vignettes). A total of twenty cognitive interviews were conducted over a span of about 8-10 weeks; most of these interviews were administered over the telephone. After each round of testing, the design team met to discuss findings and make modifications to the draft questionnaire. As alluded to above, a variety of constraints were imposed on the design and evaluation process: (a) a tight timeline for questionnaire development; (b) limited resources for both presurvey and postsurvey evaluation work; (c) a questionnaire with a strictly limited set of items (i.e., to minimize burden and cost); and (d) limited degrees-of-freedom with respect to the wording used in certain questionnaire items.

These constraints notwithstanding, pretesting work detected (and endeavored to correct) a variety of problems with the draft questionnaire. For example, with respect to Q1, an effort was made to clarify what was meant by a “landline (fixed-line) telephone”; and with respect to Q3, an effort was made to improve the list of response options. As a result, the design team was confident that the final version of the supplement questionnaire (Table 1, appendix) was a distinct improvement over the initial draft (Table 2, appendix). Subsequent to cognitive testing, and prior to the administration of the supplement in February 2004, a small-scale operational field test (about 600 CATI cases) was conducted by the Census Bureau to determine if the instrument worked as intended. To my knowledge, no substantive evaluation of the performance of the supplement questionnaire was conducted by BLS staff during this operational field test.

Postsurvey research involved the use of two evaluation methods: behavior coding and interviewer debriefing. Behavior coding was conducted at two telephone centers during the first three days of CPS interview week (15-17 February 2004). Initial coding was done *on-line*, that is, while interviews were in progress. A survey methodologist (the present author) monitored CPS interviews, selected cases that had not yet advanced to the supplement stage, and coded exchanges that took place between interviewers and respondents during administration of the supplement. For each supplement item, a maximum of two behavior codes on either side of a particular interviewer-respondent exchange were recorded (see Table 3, appendix, for a listing of interviewer and respondent behavior codes). While an effort was made to code all of the item-specific exchanges that took place between interviewers and respondents—a difficult task when

^{*} Regarding my role and responsibilities in this effort, I was asked by one of the sponsor’s representatives to conduct the postsurvey evaluation work (behavior coding and interviewer debriefing); however, prior to conducting that work, I was also provided with the opportunity to monitor many of the cognitive interviews that were conducted during the presurvey evaluation phase. On that basis, I made a number of suggestions to the design team regarding item wording; some of those suggestions were adopted, others were not.

coding is conducted on-line—only data for the first interviewer-respondent exchange have been included in coding tabulations. In all, behavior coding data were collected for 60 households. With regard to interviewer debriefing, evaluative information and data were gathered using a focus group format. During the focus-group sessions, quantitative data were collected using a *rating form* (i.e., for assessing the response difficulty of those items spontaneously identified as problematic); qualitative information was collected using a protocol of *scripted probe questions* (i.e., for gathering information on the nature of item-specific problems) and a set of *ad hoc probe questions* (i.e., for assessing the degree to which interviewers understood the objectives of supplement item Q3).

3.4. Supplement Metadata

As is the case for all CPS supplements, the sponsors drafted an instructional memorandum for interviewers several months prior to the supplement's administration date (US Bureau of the Census, 2004). Instructional memoranda provide information on the purpose of the supplement, item objectives, key definitions and other information that might be useful to interviewers in conducting the survey. Depending on the length of the supplement questionnaire, guidance is not always provided for every questionnaire item; classification items typically receive the most attention in these memoranda.

3.5. Findings from Evaluation Research

To my knowledge, no formal reports were written documenting the three rounds of cognitive interviewing, though summaries were prepared and distributed after each round for the benefit of the design team. However, formal reports were written documenting postsurvey evaluation research, and some of the information/data contained in those reports is reproduced here (see Tables 3 through 5, appendix). To simplify the presentation of findings, the information/data provided on subsequent pages will focus on three supplement items: Q1, Q2 and Q3. The first two items, Q1 and Q2, are central to algorithms used to generate supplement estimates; *Q3 data are not used in any of the estimation algorithms.*

Item Q1. The *objective* of Q1 is to obtain an accurate count of the number of distinct landline telephone numbers that provide service to the sample household. However, not all of the lines reported by respondents are used for incoming person-to-person calls (e.g., some are used for fax machines or computers); subsequent items (Q1a and Q1b) gather data on actual usage. Among other issues, the cognitive interviews led us to expect possible problems with the response task (e.g., confusion with respect to reporting distinct *telephone numbers* versus the number of telephones in the household) and with the intended meaning of technical terms (e.g., “fixed line telephone number”; “landline telephone number”). For example, the term “*landline* telephone number” was unfamiliar to some research participants (especially older persons) and seemed unnatural to others; at least fourteen alternative ways of communicating about a landline telephone were mentioned spontaneously during the cognitive interviews (e.g., home phone; house phone; our regular telephone number; main line; regular phone line.) To address the response-task problem, a second verification item (“VER2”) was incorporated into the questionnaire for responses of “two or more” to clarify question intent and ensure respondents were reporting distinct telephone numbers and not the number of telephones in the household. To address the terminology problem, the term “landline telephone number” was specifically defined in the final version of this item and an extended discussion of this technical concept was provided in the supplement instructional memorandum. In spite of these efforts, interviewers and respondents still struggled with Q1. With regard to behavior coding data (see Table 3, appendix),

interviewers read the question as worded 62% of the time; there were major changes in question wording 22% of the time—in most cases, definitional material was omitted. Respondents provided adequate (though not necessarily accurate) answers 95% of the time, but felt the need to elaborate on their answers in 15% of the cases. With regard to interviewer debriefing data (see Table 4, appendix), this item was rated eighth (of twelve) in terms of difficulty. Some of the problems identified during pretesting were not completely resolved. For example, some elderly respondents were still having issues with the term “landline” and one such respondent actually started counting the number of digits in her telephone number (a total of ten) rather than the number of landline phones in her household with distinct telephone numbers. These problems notwithstanding, two verification items (“VER1” and “VER2”) no doubt play a significant role in minimizing the level of measurement error associated with Q1.

Item Q2. The *objective* of Q2 is to determine if anyone in the sample household (*excluding* students who may be living away at school) owns a cellular telephone with a working number. Among other issues, the cognitive interviews led us to expect possible problems with the response task (e.g., whether to include/exclude household members who were living away at school) and with the intended meaning of technical terms (e.g., “*working* cell phone number”). To address the response-task problem, a phrase was inserted at the front of Q2 instructing respondents to exclude students living away at school. To address the terminology problem, the term “*working* cell phone number” was specifically defined in the supplement instructional memorandum and an extended discussion of this technical concept—including a chart classifying various types of cutting-edge communication devices (e.g., “blackberries”)—was provided in the memorandum as well. Though no doubt successful in precluding many of the more serious problems that might have arisen during supplement administration, these efforts did not resolve all of the issues associated with Q2. With regard to behavior coding data, interviewers read the question as worded 90% of the time, and there were relatively few cases (5%) where major changes in question wording were observed. Respondents provided adequate (though not necessarily accurate) answers 97% of the time, but felt the need to elaborate on their answers in 20% of the cases. In almost every recorded instance of elaboration, it appeared that respondents were simply trying to be informative when offering their response (e.g., “Yes, my wife and I both have one.”). In one case, the respondent answered “yes” but quickly added that she did not want to give out those numbers. With regard to interviewer debriefing data, this item was rated ninth (of twelve) in terms of difficulty. Some of the problems identified during pretesting were not completely resolved. For example, some respondents were not sure whether their prepaid cell phones counted as a working cell phone number. The prepaid-phone issue was addressed in the instructional memorandum (i.e., yes, they do count); however, such information will be of little use to respondents if they are not motivated to ask the interviewer for clarification when Q2 is posed. Another problem noted by interviewers, but not specifically identified during the cognitive interviews, was whether to count cell phones that were provided by an employer (and used primarily for work) as a “working” number.

Item Q3. The *objective* of Q3 is “to determine if the household relies most heavily on the cell phone number.” Though not specifically mentioned in the body of the question (but addressed in the instructional memorandum), the reference period for this item was specified as “a typical week.” Not specified in the memorandum were the following: (a) to whom in the household this question pertains (e.g., *everyone—including children—on the household roster*; just adults; just persons who own a cell phone); and (b) to which types of calls does this question pertain (e.g., *calls received anywhere*; just calls received at home). [Note: The correct answers regarding these

two interpretations appear in italics above.]^{*} Among other issues, the cognitive interviews led us to expect possible problems with the response task (e.g., should respondents consider all calls received, both at home and away from home, or just calls received at home), with the intent of the question (e.g., whether the sponsors are interested in counting all calls received—even if screened via “Caller ID” and never actually answered—or only those calls actually answered at the time they were received), and with the meaning of technical terms (e.g., “*all* of the phone calls”; “receive/received”). Other than changing the set of response options, making a few minor wording changes in the body of the question and defining the reference period in the instructional memorandum, no other steps were taken to address the concerns raised above. With regard to behavior coding data, interviewers read the question as worded 73% of the time and with minor wording changes 23% of the time. Most of the minor changes involved the response precodes, adding/deleting a word. Respondents clearly struggled to provide adequate answers to this item (63%). One out of every two responses was initially problematic in some respect: 13% inadequate answers; 13% requests for clarification; and 17% “other” responses. A response of “half” accounted for most of the inadequate answers. With regard to interviewer debriefing data, this item was rated first in terms of difficulty. And, as one might have suspected, problems identified during pretesting were not completely resolved. For example, some respondents remained uncertain as to the response task. (e.g., what household members to include in the calculations; reference period). Some of the respondents reporting for large households struggled with the estimation task; others appeared to invest very little time or effort in generating an answer to a question that should have required a series of potentially difficult mental calculations (i.e., apparent satisficing behavior). Lastly, interviewers complained repeatedly about the item’s incomplete response scale (i.e., no “half” option), noting that some respondents were adamant about that being their answer.

4. Discussion and Closing Remarks

One does not have to have twenty-five years of survey experience to recognize that the design and evaluation of the CPS supplement described above was not optimal. Much more work could have been undertaken in the following areas:

- Conceptually (with respect to design), more could have been done in early developmental stages to understand how families and individuals use telephones, cell phones and other communication devices (e.g., focus groups with families and/or industry representatives; see Gower and Haarsma, 1997).
- Methodologically (with regard to presurvey evaluation work), more could have been done after the three rounds of cognitive interviews to determine how the draft questionnaire would work in a field setting (i.e., a small-scale field test that focused on the questionnaire and not simply on operational aspects of the instrument).
- Pragmatically (with respect to design), more could have been done to implement design changes (and upgrade supplement metadata) based on the information gathered during the cognitive interviews—especially with respect to Q3.

^{*} During the process of behavior coding, it became obvious that some respondents were having difficulty with the estimation task imposed by Q3. In an effort to determine what interviewers understood the intent of this question to be, I decided to ask a set of unscripted debriefing questions during the two focus groups that followed several days later. Recall that only the reference period was specified in the supplement instructional memorandum. The answers to those unscripted debriefing questions are summarized in Table 5, appendix. Surprisingly, interviewers were least accurate in their responses on the only element that was explicitly specified (i.e., the reference period).

- Methodologically (with regard to postsurvey evaluation work), more could have been done to obtain quantitative estimates of measurement error (e.g., carefully crafted respondent-debriefing items).*

That said, and given what was done and *not done* in the research described above, what lessons can we take away from this case study and how might such a study guide our behavior as survey practitioners? From my perspective, the primary lessons are these: First, resources for questionnaire design-and-evaluation work are often limited (and may become more so in the future); the greater the resource constraints, the less likely it is that practitioners will have the means to make good design decisions and conduct credible evaluation research. And secondly, given that we in the QUEST community have developed a rich understanding of how to conduct survey-design-and-evaluation research *well* (and the potential consequences of *not doing this work well*), difficult professional decisions inevitably will need to be made with regard to participation in low-resource-type research efforts.

These lessons have implications for professional behavior, of course, and my advice to myself and to other practitioners who might be interested in such advice would be this, “Primum non nocere”: First [and foremost], do no harm. Well, what might that mean exactly? In my view, it means that practitioners should seek to minimize the potential for survey-related error by making every effort to adhere to the highest standards established by their profession (see Reference section). It also means possibly walking away from a specific design-and-evaluation research effort if, after making one’s case to survey sponsors, a survey methodologist strongly suspects that those standards are likely to be compromised. Integrity is paramount in our profession (indeed, in all professions); competence, though obviously important, must be viewed as secondary to this essential attribute. Taking this case study as an example, I believe that professional integrity would require a full documentation of the design-and-evaluation process, its constraints and its findings—whatever the consequences might be. It is worth noting that this oath “to do no harm” (as it applies to the survey methodology domain), not only safeguards the credibility of practitioners, but also the credibility of the organizations we serve and the myriad professionals who rely on the quality of our data to make policy decisions.

Let me close (and summarize) with two assertions for your consideration: If the resources available for a particular research undertaking are limited such that we are not capable of doing

* For example, given the prevalence of call-screening devices, one could reasonably assume that some respondents might not want to report that they take incoming calls on a landline number—the motive being to avoid receiving/taking calls from unfamiliar parties. One indirect means of testing such a hypothesis would be to review response-distribution data and analyze cross-tabulation data to uncover highly unlikely response patterns—and I did so by examining supplement items **Q1b** and **Q2**. When items Q1b and Q2 were cross-tabulated (total N=5940), approximately 10% (n=570) of the respondents who said they did not have a cell phone (Q2: “no”) also said they did not take incoming calls on their only landline number (Q1b: “no”). Now, given the high cost of having a landline number—and not owning a cell phone or any other obvious means of communication with the outside world—why would respondents say that they do *not* take incoming calls on their only landline number? There are plausible reasons, to be sure (e.g., no friends or family; only communicate via computer); however, it seems more likely that a fair number of respondents may simply wish to avoid being contacted by individuals who conduct surveys or sell unwanted products or services—and if so, they may misreport. The point of this illustration is that this issue (and other logical inconsistencies in the data) could have been addressed by developing a set of response-specific debriefing questions for just this sort of situation. The 570 respondents in this group could have been asked the following open-ended debriefing question: “You mentioned earlier that you do not take incoming calls on your landline number. If there were an emergency involving friends or family, by what means could a concerned individual contact you?” The data/information provided by asking such a question, not to mention the response latency, could potentially be very useful. [Esposito, 2004(b), p. 21]

whatever research needs to be done in a professionally acceptable manner, then the most prudent course of action may be not to participate at all. Should we feel compelled to participate, a plan for thorough documentation of all aspects/phases of the design-and-evaluation process should be discussed with sponsors before research commences and formalized in writing.

References

- Akkerboom, H., and Dehue, F. (1997). The Dutch Model of Data Collection Development for Official Surveys. *International Journal of Public Opinion Research*, 9, pp. 126-145.
- Beatty, P. (2004). "The Dynamics of Cognitive Interviewing." In S. Presser, J. Rothgeb, et al. (eds.) *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley-Interscience, pp. 45-66.
- Beukenhorst, D., Giesen, D., deVree M. (2001). "Computerized versus Interviewer-Guided Evaluation of CASI Questionnaires." QUEST2001: Proceedings of the Third [Workshop] on Questionnaire Evaluation Standards. Washington, DC: U.S. Bureau of the Census, pp. 57-63.
- Cosenza, C., and Fowler Jr., F.J. (2001). "Learning from Cognitive Interviews: Fact or Fiction?" QUEST2001: Proceedings of the Third [Workshop] on Questionnaire Evaluation Standards. Washington, DC: U.S. Bureau of the Census, pp. 52-56.
- 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.
- Esposito, J.L. (2004a). "Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study." *Journal of Official Statistics*, 20, pp.143-183.
- Esposito, J.L. (2004b). "An Evaluation of the CPS Cell-Phone-Use Supplement: Composite Report." Unpublished final report. Washington, DC: Bureau of Labor Statistics.
- 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.
- 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 Jr., F.J. (2004). "The Case for More Split-Ballot 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.
- Gower, A.R., and Haarsma, G. (1997). "A Comparison of Two Methods in a Test of the Canadian Census Questionnaire: Think-Aloud Focus Groups vs. Focus Groups." Paper presented at QUEST1997, the First [Workshop] on Questionnaire Evaluation Standards. Örebro, Sweden. [Originally called the "MIST" Workshop: Minimum Standards in Questionnaire Testing.]
- Haraldsen, G. (2004). "Identifying and Reducing Response Burdens in Internet Business Surveys." *Journal of Official Statistics*, 20, pp. 393-410.
- Lindström, H., Davidsson, G., Henningsson, B., Björnram, A., Marklund, H., Denell, C., Hoff, S. (2004 English/2001 Swedish). *Design Your Questions Right: How to Develop, Test, Evaluate and Improve Questionnaires*. Örebro, Sweden: Statistics Sweden.
- 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.

- Potaka, L., and Cochrane, S. (2004). "Developing Bilingual Questionnaires: Experiences from New Zealand in the Development of the 2001 Māori Language Survey." *Journal of Official Statistics*, 20, pp. 289-300.
- 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.
- Prüfer, P., Rexroth, M., and Fowler Jr., F.J. (Eds.) (2004). *QUEST2003: Proceedings of the Fourth [Workshop] on Questionnaire Evaluation Standards*. Spezial Band 9. Mannheim, Germany: ZUMA.
- Prüfer, P., and Rexroth, M. (1999). "Using Cognitive Techniques in the Field." QUEST1999: Proceedings of the Second [Workshop] on Questionnaire Evaluation Standards. London, England: Office of National Statistics, pp. 95-99.
- 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.
- Snijkers, G. (2002). "Cognitive Laboratory Experience: On Pretesting Computerized Questionnaires and Data Quality." Ph.D. Dissertation. Utrecht University, Utrecht, and Statistics Netherlands, Heerlen.
- 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.]
- Tucker, C., Brick, J.M., Meekins, B., and Morganstein, D. (2004). "Household Telephone Service and Usage Patterns in the U.S. in 2004." Draft of paper presented at the 2004 Joint Statistical Meeting of the American Statistical Association.
- U.S. Bureau of the Census (2004). Interviewer Manual for the Cell Phone Use Supplement to the Current Population Survey. CPS Field Representative/CATI Interviewer Memorandum Number 2004-02 [Field Division]. Washington, DC: U.S. Department of Commerce.
- Willimack, D.K., Lyberg, L., Martin, J., Japac, 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.

Table 1. CPS Cell-Phone-Use Supplement: Final Question Wording

Label	Final Supplement Question Wording [February 2005]
Q1	<p>First I would like to ask about any regular, landline telephone numbers in your household. These numbers are for phones plugged into the wall of your home and they can be used for different reasons, including making or receiving calls, for computer lines or for a fax machine.</p> <p>How many different landline telephone numbers does your household have?</p>
VER1	<p>I'd like to verify the information you just provided. I believe you indicated that your household has NO LANDLINE TELEPHONE service for incoming and outgoing calls: Is that correct?</p>
VER2	<p>I just want to verify that your household has [fill Q1] distinct telephone NUMBERS: Is that correct?</p>
Q1a	<p>Excluding any numbers used only for faxes and computers, how many of these [fill Q1] landline telephone numbers are used for incoming calls?</p>
Q1b	<p>Excluding a number used only for a fax or computer, do you [fill (or any other members of your household) if NUMHOU > 1] take incoming calls on a landline number?</p>
Q2	<p>[Fill (Excluding students living away at school,) if NUMHOU>1] Do you [fill (or any other members of your household) if NUMHOU > 1] have a working cell phone number?</p>
Q2a	<p>[Fill (Excluding students living away at school,) if NUMHOU>1] How many different cell phone numbers [fill (do you have?) if NUMHOU = 1 or fill (do the members of your household have?) if NUMHOU (number of persons in household) >1]</p>
Q2b	<p>How many of the [fill Q2a] cell phone numbers you have do you [fill (or any other members of your household) if NUMHOU > 1] use regularly?</p>
Q2c	<p>How many of the [fill Q2a] cell phone numbers are answered by more than one household member?</p>
Q2d	<p>Do you [fill (or members of your household) if NUMHOU > 1] regularly answer this cell phone number?</p>
Q2e	<p>Is this cell phone number answered by more than one household member?</p>
Q3	<p>Of all the phone calls that you [fill (or any other members of your household) if NUMHOU > 1] receive, about how many are received on a cell phone? Would you say ...</p> <p><1> All or almost all calls, <2> More than half, <3> Less than half, or <4> Very few or none?</p>

Table 2. Early Draft Question Wordings for Selected Items [Q1, Q2 and Q3].

Label	Question Wording [Early draft of items, circa May 2005]
Q1	How many different fixed line telephone numbers will reach your household? [VERIFY ZERO and SKIP to Q2: “May I please verify that you do not have any regular fixed line telephone numbers in your home—by this I mean the type of telephone numbers homes with telephones had before cell phones were available.”]
Q2	Do you or any other members of your household have a working cellular phone? [IF DK: “Please remember that all of the information you are providing is confidential.”]
Q3	Of all the incoming calls this household takes, how many are received at home on a cell phone? Would you say: (1) All (2) Most (3) Some (4) Hardly any, or (5) None (97) REF (98) DK

Table 3. Percentage and Frequency of Interviewer and Respondent Behavior Codes for Twelve Supplement Items

Q Label	Interviewer Codes ¹				Respondent Codes ¹							Comments ²		
	N	E	mC	MC	PVF	N	AA	qA	IA	RC	INT		D/R	O
Q1	(60)	62% (37)	17% (10)	22% (13)	3% (2)	(60)	95% (57)		3% (2)	2% (1)	7% (4)		15% (9)	PVF: P, F
VER1	(2)	100% (2)				(2)	100% (2)							Low N.
VER2	(10)	80% (8)	10% (1)	10% (1)		(10)	90% (9)	10% (1)					20% (2)	
Q1a	(12)	100% (12)			17% (2)	(12)	83% (10)		8% (1)	8% (1)			8% (1)	PVF: P, V
Q1b	(8)	75% (6)		25% (2)		(8)	100% (8)							Low N. Data are an artifact of entry errors (see section II.B.)
Q2	(58)	90% (52)	5% (3)	5% (3)	5% (3)	(59)	97% (57)		2% (1)	2% (1)	2% (1)		20% (12)	PVF: P-, V, V
Q2a	(40)	85% (34)	8% (3)	5% (2)	5% (2)	(40)	100% (40)				3% (1)		3% (1)	PVF: P, V
Q2b	(23)	44% (10)	52% (12)	4% (1)	13% (3)	(23)	96% (22)		4% (1)				17% (4)	PVF: P, P-, P-
		Continued on Next Page												

Superscript 1: Because of multiple codes being assigned for a particular question, row percentages for interviewer or respondent behavior codes may sum to values greater than 100 percent.
Superscript 2: In the "Comments" column, entries to the left of the colon refer to a particular column in the table (e.g., PVF) and values to the right indicate what the actual observations enumerated in that column were (e.g., "V,Vs" refers to one regular verify code and one silent verify code).

ABBREVIATIONS: "N" refers to the number of times a question was asked (interviewer behavior codes) or a response given (respondent behavior codes); N is the base for all percentage calculations in a particular row. With regard to interviewer codes: "E" refers to an exact question reading, "mC" to a minor change in question wording, "MC" to a major change in wording, and "PVF" to probe, verify, or feedback, respectively. "Vs" refers to a silent verify (i.e., interviewer enters information the respondent provided earlier in lieu of asking the question). With regard to respondent codes: "AA" refers to an adequate answer (i.e., an answer that matches a precoded response category), "qA" refers to a qualified answer, "IA" refers to an inadequate answer (i.e., one that does not match a precoded response category), "RC" refers to a request for clarification, "INT" refers to an interruption (usually with an answer) by the respondent, "D" refers to a response of "don't know", "R" refers to a refusal to answer the question, and "O" refers to other (i.e., a miscellaneous category). Use of the negative sign (-) indicates that a particular interviewer behavior was poorly executed; for example, V- might refer to a probe question that was leading.

Table 3. (continued)

Q Label	Interviewer Codes ¹			Respondent Codes ¹							Comments ²			
	N	E	mC	MC	PVF	N	AA	qA	IA	RC		INT	D/R	O
Q2c	(21)	95% (20)			5% (1)	(20)	85% (17)		10% (2)	5% (1)			15% (3)	PVF: P-
Q2d	(17)	88% (15)	6% (1)	6% (1)	6% (1)	(16)	94% (15)		6% (1)				19% (3)	PVF: F
Q2e	(9)	78% (7)		22% (2)		(9)	89% (8)		11% (1)				11% (1)	Low N.
Q3	(30)	73% (22)	23% (7)	3% (1)		(30)	63% (19)	3% (1)	13% (4)	13% (4)	3% (1)		17% (5)	

Superscript 1: Because of multiple codes being assigned for a particular question, row percentages for interviewer or respondent behavior codes may sum to values greater than 100 percent.

Superscript 2: In the "Comments" column, entries to the left of the colon refer to a particular column in the table (e.g., PVF) and values to the right indicate what the actual observations enumerated in that column were (e.g., "V.V.s" refers to one regular verify code and one silent verify code).

ABBREVIATIONS: "N" refers to the number of times a question was asked (interviewer behavior codes) or a response given (respondent behavior codes); N is the base for all percentage calculations in a particular row. With regard to interviewer codes: "E" refers to an exact question reading, "mC" to a minor change in question wording, "MC" to a major change in wording, and "PVF" to probe, verify, or feedback, respectively. "Vs" refers to a silent verify (i.e., interviewer enters information the respondent provided earlier in lieu of asking the question). With regard to respondent codes: "AA" refers to an adequate answer (i.e., an answer that matches a precoded response category), "qA" refers to a qualified answer, "IA" refers to an inadequate answer (i.e., one that does not match a precoded response category), "RC" refers to a request for clarification, "INT" refers to an interruption (usually with an answer) by the respondent, "D" refers to a response of "don't know", "R" refers to a refusal to answer the question, and "O" refers to other (i.e., a miscellaneous category). Use of the negative sign (-) indicates that a particular interviewer behavior was poorly executed; for example, V- might refer to a probe question that was leading.

Table 4. Difficulty Ratings Assigned to Problematic Supplement Items

Item	TC	Interviewer Number										Mean	SD
		1	2	3	4	5	6	7	8	9	10		
<i>Q1</i>	TTC	2	1	2.5	1	3	1	1	2	1	1	1.55	0.762
	HTC	1	5	1	2	2	2	1	1	2	3	2.00	1.247
Totals:											1.78	1.032	
<i>VER1</i>	TTC	<i>io</i>	<i>io</i>	<i>io</i>	2	2	<i>io</i>	1	<i>io</i>	<i>io</i>	1	1.50	0.577
	HTC	-	-	-	-	-	-	-	-	-	-	-	-
Totals:											1.50	0.577	
<i>VER2</i>	TTC	<i>io</i>	<i>io</i>	4	4	1	1	3	1	1	1	2.00	1.414
	HTC	1	1	<i>io</i>	1	2	1	1	1	1	1	1.11	0.333
Totals:											1.53	1.068	
<i>Q1a</i>	TTC	-	-	-	-	-	-	-	-	-	-	-	-
	HTC	1	2	<i>io</i>	1	1	1	1	1	1	1	1.11	0.333
Totals:											1.11	0.333	
<i>Q1b</i>	TTC	<i>io</i>	<i>io</i>	5	2	4	2	2	2	3	3	2.88	1.126
	HTC	2	3	<i>io</i>	1	2	<i>io</i>	2	1	1	1	1.63	0.744
Totals:											2.25	1.236	
<i>Q2</i>	TTC	2	1	3.5	2	2	<i>b</i>	2	1	1	4	2.06	1.074
	HTC	1	3	1	1	1	2	2	1	1	2	1.50	0.707
Totals:											1.76	0.919	
<i>Q2a</i>	TTC	2	1	3	2	2	1	2	1	2	4	2.00	0.943
	HTC	-	-	-	-	-	-	-	-	-	-	-	-
Totals:											2.00	0.943	
<i>Q2b</i>	TTC	2	2	4.5	3	4	1	3	1	5	2	2.75	1.399
	HTC	1	3	3	1	2	1	5	1	1	2	2.00	1.333
Totals:											2.38	1.385	
<i>Q2c</i>	TTC	3	1	5	2	3	3	2	3	3	5	3.00	1.247
	HTC	1	2	3	3	3	2	4	1	1	1	2.10	1.101
Totals:											2.55	1.234	
<i>Q2d</i>	TTC	-	-	-	-	-	-	-	-	-	-	-	-
	HTC	1	3	1	3	2	1	3	2	1	1	1.80	0.919
Totals:											1.80	0.919	
<i>Q2e</i>	TTC	-	-	-	-	-	-	-	-	-	-	-	-
	HTC	1	4	4	2	3	1	3	3	1	1	2.30	1.252
Totals:											2.30	1.252	
<i>Q3</i>	TTC	2	1	5	3	4	1	4	2	2	3	2.70	1.337
	HTC	1	4	2	4	4	1	4	2	2	1	2.50	1.354
Totals:											2.60	1.314	

Table 4 continues on the next page.

Table 4. continued

<p>Question and Scale Used to Rate Problematic Supplement Items:</p> <p>Q. Based on your experiences this past week, about how frequently did the <i>respondents</i> you interviewed have difficulty providing an adequate answer to this question?</p> <ul style="list-style-type: none"> ▪ A/1: Never or rarely → 0 to 10% of the time ▪ B/2: Occasionally → some % between A and C ▪ C/3: About Half the Time → approximately 40-to-60% of the time ▪ D/4: A Good Deal of the Time → some % between C and E ▪ E/5: Almost Always or Always → 90 to 100% of the time <p>Abbreviations: “TC” for telephone center; “TTC” for Tucson Telephone Center; “HTC” for Hagerstown Telephone Center; “b” for blank entry; “io” for insufficient observations to rate item.</p> <p>Notes: TTC interviewer number 3 assigned two precodes to several items which resulted in fractional (average) values for these items. Dashes (-) signify that the item was not identified as problematic by a group of interviewers and therefore was not rated.</p>

Table 5. Four Debriefing Questions Targeting Supplement Item Q3

Total N TTC+HTC=n	Debriefing Questions
N=20 8+8=16 1+1=2 1+1=2 0+0=0	<p>DQ1: To whom in the household does Q3 pertain?</p> <p><a> <i>Everyone listed on the household roster (adults and children)</i></p> <p> Just adults and older children (15+)</p> <p><c> Anyone in the household who owned a cell phone</p> <p><d> Other</p>
N=19 6+8=14 0+0=0 3+1=4 0+1=1	<p>DQ2: To which types of telephone calls does Q3 pertain?</p> <p><a> <i>All landline and cell phone calls received at home, work, shopping, etc.</i></p> <p> To landline and cell phone calls received at home and work only</p> <p><c> To landline and cell phone calls received at home only</p> <p><d> Other</p>
N=20 3+1=4 1+6=7 2+1=3 4+2=6	<p>DQ3: What do you think the reference period might be for Q3?</p> <p><a> Typical month</p> <p> <i>Typical week</i></p> <p><c> Typical day</p> <p><d> Other</p>
N=20 4+8=12 6+1=7 0+1=1	<p>DQ4: Did respondents understand Q3 the same way you did?</p> <p><a> Yes</p> <p> No</p> <p><d> Other</p>
<p>Note: <i>Correct answers</i>, as specified by the sponsor (and/or the supplement interviewer manual) appear in italics for DQ1, DQ2 and DQ3. Also, in reading these hand-written questions to interviewers, the moderator embellished question wording in an effort to enhance comprehension.</p>	