Sixth International Conference on Social Science Methodology [RC33: Logic and Methodology], Amsterdam, The Netherlands; 16-20 August 2004

Looking for Trouble and Finding It, Sometimes: Exploring Relationships between Pre-survey and Post-survey Evaluation Data

James L. Esposito, Bureau of Labor Statistics (USA) [1]

Key words: Behavior coding, cognitive interviews, CPS cell-phone-use supplement, focus groups, interviewer debriefing, questionnaire appraisal system.

Abstract: Using a new Current Population Survey (CPS) supplement on landline and cell-phone service and use as the investigative context, this case study explores relationships between pre-survey evaluation data (drawn from cognitive interviews and a questionnaire appraisal coding system) and post-survey evaluation data (drawn from behavior coding and interviewer debriefings). Using qualitative data from cognitive interviews and the questionnaire appraisal system (Willis and Lessler, 1999), predictions were formulated as to where problems with the supplement questionnaire might occur during its administration in February 2004. Evidence of problems was based on behavior-coding data from 60 household interviews and on qualitative data from two focus groups conducted with CPS interviewers. Though subjective (i.e., no means of quantifying measurement error was available), the accuracy of predictions was assessed using post-survey evaluation data. A summary of predictive "hits" and "misses" is provided and discussed within the context of a larger questionnaire-design-and-evaluation framework (Esposito, 2004a, 2004b) that relates pre-survey and post-survey evaluation work.

I. Introduction

To maximize the public's return on investments intended to advance the common good, statistical survey research organizations are expected to allocate scarce resources wisely in the process of accomplishing their stated objectives/missions. Given limited resources (i.e., staff and money), this often leads to difficult choices with respect to survey design and evaluation work. Tradeoffs between survey errors—one aspect of data quality—and survey costs inevitably come into play (see Groves, 1989). Given the reality of such tradeoffs, what options are available to a survey sponsor when a prospective survey is not mission-critical (in the substantive sense), when resources are limited, and when turn-around time for design-and-evaluation work is unusually tight? Two options come to mind: (1) the sponsoring organization can do the best it can with the limited resources

[1] Bureau of Labor Statistics, Postal Square Building, Room 4985, 2 Massachusetts Avenue, N.E., Washington, DC, 20212, USA (Esposito.Jim@bls.gov). *The views expressed in this paper are those of the author and do not reflect the policies of the Bureau of Labor Statistics (USA).* I wish to acknowledge the following individuals: Clyde Tucker for providing subject-matter expertise; David Cantor and Tracey Haggerty-Heller for graciously sharing data they collected during the pretesting phase of this research; and Lisa Clement and Dennis Clark for their technical assistance. I also wish to thank the interviewers and managers at the Census Bureau's Hagerstown and Tucson telephone centers who participated in this project, and the SORT team. This paper could not have been written without their significant contributions.

and time it has available, or (2) it can choose not to conduct the survey until conditions become more favorable. This paper is about some of the ramifications associated with option one, especially with respect to the evaluation components of the questionnaire design-and-evaluation process.

I.A. The Cell-Phone-Use Supplement: Rationale and Objectives

The investigative context for this research was the development and implementation of a new Current Population Survey (CPS) supplement on landline and cell-phone service and use. The *cell-phone-use supplement* (see **Table 1**), as it came to be known, 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 was not being reached—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 survey organizations are having more and more trouble reaching landline-telephone households. It was hoped that the supplement would provide information on patterns of telephone usage in American households, especially with respect to how households with both landlines and cell phones use the two technologies. The first draft of the supplement questionnaire was developed by a group of subject-matter experts from government, academia, and the private sector. [2]

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 (Memorandum from Esposito to Tucker, 31 July 2004).

[2] The substantive information provided in this paragraph draws on an e-mail from Clyde Tucker to Jim Esposito (06 May 2004).

I.B. Division of Work

As noted, the initial draft of the supplement questionnaire was developed by a group of subject-matter experts from government, academia, and the private sector. Later drafts were refined by subject-matter and questionnaire-design specialists on the basis of several rounds of cognitive testing conducted by private-sector researchers. Responsibility for developing supplement metadata (e.g., interviewer instructions; classification algorithms) was to be assumed by representatives of the two supplement sponsors. In June 2003, the present author was asked to join a small research team that had developed and had commenced cognitive testing on the supplement. My role in this process was as follows: (1) to contribute to the development of supplement instructional materials for CPS interviewers, and (2) to assume responsibility for conducting a modest evaluation of the supplement when it was first administered in February 2004. However in July, after being given the opportunity to review audiotapes of cognitive interviews and to monitor some of these mostly telephone interviews while they were in progress, I was invited to participate in several teleconference calls and to provide the BLS sponsor with memos documenting observations, comments and suggestions for possible design modifications. Given time constraints and limited opportunities for subsequent testing, *some of the suggestions were adopted and others were not*. Then in December 2003, after reviewing a draft of the supplement instructions that the survey sponsors had prepared for CPS interviewers, I provided the BLS sponsor with comments and a set of recommendations for modifying the interviewer instructions. Again, given various constraints, *some of the recommendations were adopted and others were not*.

II. Research Methodology

As the title of the paper suggests, the principal objective of this study was to explore relationships between pre-survey and post-survey evaluation data. In pursuing this objective, two paths were chosen. First, using qualitative data from three sources (i.e., cognitive interviews; a system for questionnaire appraisal; an informal "expert review" of the cell-phone-use questionnaire by a select group of CPS interviewers), predictions were formulated as to where problems with supplement items might occur during its administration in February 2004. Though an admittedly a subjective process (i.e., no

means of quantifying measurement error was available), the accuracy of predictions was assessed using post-survey evaluation data drawn from behavior-coding work and interviewer debriefings. The second path taken was to compute correlations between various quantitative indicators (i.e., those generated by pre-survey and post-survey evaluation methods) as a means of exploring whether relationships between these methods existed. These objectives have the effect of binding this research to a family of other studies with similar objectives and goals (e.g., Forsyth, Rothgeb and Willis, 2004; Hess, Singer and Bushery, 1999; Presser and Blair, 1994; Rothgeb, Willis and Forsyth, 2001; Willis, Schechter and Whitaker, 1999). However, the pragmatic and exploratory nature of this research may make it difficult for some readers to notice the family resemblance. [3]

II.A. Pre-Survey Evaluation Work.

As mentioned above, the pre-survey evaluation work included cognitive interviews, an informal expert review, and use of a questionnaire appraisal system. The various evaluation methods are described below.

II.A.1. *Cognitive Interviews.* Over the past dozen years or so, in recognition of the central (but not solitary) role that cognition plays in the survey response process, the method of cognitive interviewing has become an essential component of questionnaire design and evaluation work and, quite appropriately, this method has received a great deal of attention in the survey methodology literature (e.g., DeMaio and Landreth, 2004; Gerber, 1999; Gerber and Wellens, 1997; Willis, 2004; Willis, DeMaio and Harris-Kojetin, 1999). Cognitive interviewing has been found to be especially useful in identifying problems that respondents experience in comprehending the intent and conceptual content of survey questions. During June and July (2003), three rounds of cognitive interviews were conducted to evaluate and refine draft versions of the cell-phone-use supplement. A total of twenty-two cognitive interviews were conducted by two experienced behavioral scientists who work at a large, private-sector survey research organization. The interviewers developed and made use of protocols comprising items

[3] For an interesting discussion of the differences among exploratory, confirmatory and reparatory research, with respect to the utility of various survey evaluation methods, the reader had two very good options: Forsyth, Rothgeb and Willis, 2004 (pp. 526-527) and Willis, DeMaio and Harris-Kojetin (1999).

from the cell-phone-use supplement questionnaire, scripted probes, and embedded vignettes; they also asked unscripted probes when doing so proved advantageous. The interviews were administered over the telephone and audiotaped with the permission of research participants. After each round of testing, the interviewers prepared a memorandum summarizing their findings, forwarded that document to the BLS sponsor and subsequently discussed their findings with the sponsor via conference call. Modifications to the then-current draft questionnaire (and to the protocol used in the second and third rounds of testing) were made during and after the conference call. Although informative and very useful in making revisions to the draft questionnaires, the summary memos mentioned above were not utilized by the present author in the methodological research described herein. Instead, audiotapes of the cognitive interviews were obtained and carefully reviewed. Substantive parts of these interviews were transcribed and personal observations and editorial notes were inserted for subsequent analytical review. These partial transcriptions, observations and notes served two purposes: (1) they provided much of the empirical substance for preparing subsequent review-and-recommendation memos to the BLS sponsor (e.g., Esposito to Tucker, 31 July 2003); and (2) they were later used, along with other information/data, to formulate predictions as to where problems were likely to arise during the administration of the supplement in February 2004.

II.A.2. *Questionnaire Review by Survey Operations Review Team.* In September 2003, a near-final draft of the supplement questionnaire was distributed to members of the Census Bureau's Survey Operations Review Team [**SORT**] for their review and comment. The SORT team represents a very experienced group of CPS interviewers with whom internal and external program offices can consult to provide feedback on questionnaire- or interview-related issues, like informal expert reviews of draft questionnaires or the identification of problems associated with the actual administration of existing or draft survey questionnaires. In this study, members of the team were asked by BOC representatives to participate in an informal "expert review" of the twelve items on the draft supplement questionnaire and they were provided with a set of stimulus questions to structure their review. For example: "Are there any [supplement] questions that will be difficult to understand because of the telephone-specific terms used in the

question text? Are there any questions that the respondents will have difficulty answering for reasons other that the question text?" Like the information extracted from the cognitive interviews mentioned above, SORT-team comments were reviewed as part of the process of formulating predictions of where problems were likely to arise during the administration of the supplement in February 2004.

II.A.3. *The Question Appraisal System [QAS-99].* In February 2004, just prior to the week during which the cell-phone-use supplement was to be administered, the present author decided to code/evaluate the twelve supplement items using a question appraisal system (hereafter QAS) developed by Gordon Willis and Judith Lessler (1999). According to the developers of the QAS: "The questionnaire appraisal system is designed to assist questionnaire designers in evaluating survey questions, and in finding and fixing problems, before questions 'go into the field' (Willis and Lessler, 1999, p. 1-1)." In the present context, the QAS was being employed primarily to evaluate supplement items, not fix them—it was far too late for the latter; moreover, opportunities for making significant changes to the supplement questionnaire were limited from the outset. Instead, output from the QAS would be used in two ways: (1) as another source of data and information from which to formulate predictions regarding possible problems with specific supplement items; and (2) as a pre-survey evaluation method that could later be correlated with post-survey evaluation data from behavior coding (i.e., interviewer and respondent behavior codes) and from interviewer debriefings (i.e., data generated through use of a rating form for items identified as problematic). The various categories and subcategories that comprise the QAS coding form can be viewed in **Table 2**. For each of the supplement's twelve questionnaire items, a crude *quantitative indicator* was generated by simply counting the number of QAS subcategories that were coded as potentially problematic (i.e., a sum of the "yes" entries). As a final point, readers should note the following: Strictly speaking, insofar as the present author had reviewed qualitative data from cognitive interviews and from the SORT team several months prior to undertaking the QAS evaluation task, the QAS data generated here cannot be viewed as independent of the other pre-survey evaluation data that were available for this research. This may (or may not) represent a deviation from the intended use of the QAS as an evaluation tool.

II.A.4. *Formulating Predictions.* Drawing on information available from the audiotapes of twenty-two cognitive interviews (and associated review-and-recommendation memos), on feedback from the SORT team and on the QAS appraisal work, item-specific predictions were formulated as to where problems with the supplement questionnaire might occur during its administration in February 2004 (see **Table 3**). These predictions, formulated during the week prior to supplement administration, were made available to the present author's supervisor on 13 February 2004, two days prior to the start of CPS interviewing for February.

II.A.5. *Exploring Relationships between Pre-survey and Post-survey Evaluation Data.* Given expectations that certain items on the cell-phone-use supplement could prove to be problematic for survey participants, an opportunity arose to explore relationships between pre-survey evaluation data (i.e., QAS data) and post-survey evaluation data (e.g., behavior-coding data; interviewer-debriefing data; see below). Unlike the predictions described above for specific supplement items (Table 3), only very general expectations were entertained by the present author. A discussion of those expectations will be provided in subsection II.B.4, after post-survey evaluation work has been described.

II.B. Post-Survey Evaluation Work.

The bulk of the post-survey evaluation work involved conducting behavior coding and interviewer debriefings. When informative, we also made use of response distribution analyses. The various evaluation methods are described below.

II.B.1. *Behavior Coding.* Behavior coding involves a set of procedures (e.g., developing a coding form, monitoring interviews, coding interviewer-respondent exchanges, transferring coded data to a database) which have been found useful in identifying problematic questionnaire items (e.g., Esposito, Rothgeb and Campanelli, 1994; Fowler and Cannell, 1996; Morton-Williams, 1979; Oksenberg, Cannell and Kalton, 1991). The coding form used in this research incorporated *six interviewer codes* [i.e., exact question reading (E), minor change in wording (mC), major change in wording (MC), probe (P), verify (V) and feedback (F)] and *eight respondent codes* [i.e., adequate answer (AA),

qualified answer (qA), inadequate answer (IA), request for clarification (RC), interruption (Int), don't know (D), refusal (R) and other (O)].

Behavior coding was conducted at two of the Census Bureau's three telephone centers (Hagerstown, MD, and Tucson, AZ) during the first three days of CPS interview week (15-17 February 2004) and 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. A maximum of two behavior codes *on either side* of a particular interviewer-respondent exchange were recorded. While an effort was made to code all of the exchanges that took place between interviewers and respondents for each of the twelve supplement items, a difficult task when coding is conducted on-line, only data for the *first interviewer-respondent exchange* have been included in our coding tabulations. In selecting cases to code, an effort was made to avoid coding multiple cases for the same interviewer. In all, behavior-coding data were collected for 60 households. These 60 cases were administered by 52 different interviewers; to minimize the potential for bias, no interviewer was selected for coding purposes more than twice.

With regard to *interviewer codes*, previous work involving telephone-center interviewers has led us to expect very high percentages of exact question readings (i.e., E-code values at or greater than 95 percent). When the percentage of exact question readings falls below 90 percent, we flag the item as having potentially problematic wording. With regard to *respondent codes*, diagnostic procedures are not quite as straightforward (e.g., an "adequate answer" is not necessarily an *accurate* answer). While it may be comforting to find that respondents provide adequate answers over 90 percent of the time, researchers tend to focus on other codes to gain insights into the types of problems that may exist. For example, a high percentage of requests for clarification (i.e., RC-code values at or greater than 10 percent) suggests that there may be problems with a term/concept used in the target question. A high percentage of "other" responses (i.e., O-code values at or greater than 10 percent) indicates that respondents are providing more information than would be required to adequately answer a particular questionnaire item;

such behavior may indicate uncertainty as to the specific information being requested or it may reflect a desire on the part of some respondents to elaborate on their particular circumstances with respect to the question topic. Lastly, it should be noted that while behavior coding is useful in identifying problematic survey questions, it is often *necessary to use other analytical methods* (e.g., interviewer and respondent debriefings) to identify potential causes of those problems and to provide insights as to the types of modifications that could be made to improve data quality.

II.B.2 *Interviewer Debriefings*. There are a variety of ways to gather evaluative information/data from interviewers, and a substantial literature on this class of methods exists (e.g., Converse and Schuman, 1974; DeMaio, 1983; DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993). In this particular research effort, interviewers were debriefed using a focus group format. During the focus group, data were also collected using a *rating form* (i.e., for assessing response difficulty for items identified as problematic) and item-specific *debriefing questions* (i.e., for assessing interviewer understanding of supplement item Q3). Two focus groups were conducted in February 2004 at the Census Bureau's telephone centers in Tucson, AZ, and Hagerstown, MD. Instructional materials and a log form were distributed to participating interviewers well in advance of CPS interviewing in February (see **Attachments, Table A-1)**. The *log forms* were used to record any problems interviewers may have encountered with specific supplement items in the process of completing their caseloads and they were instructed to bring their forms to the debriefing sessions for reference purposes; at the end of both sessions, the moderator collected all log forms. Prior to conducting the focus groups, a debriefing plan was formulated to standardize focus group procedures and item-specific probe questions were developed to gather information on the twelve items that constitute the cell-phone-use supplement (see **Attachments, Table A-2**). Both debriefing sessions were audiotaped and written summaries were prepared from these tapes.

In general terms, the purpose of the debriefing sessions was to obtain feedback from interviewers regarding problematic aspects of the twelve supplement items. During the focus groups, participants were asked to do the following:

- to identify spontaneously any problems that they—or respondents—may have experienced when administering the cell-phone-use supplement;
- to evaluate those items identified as problematic using a rating scale provided by the moderator (see details provided in **Table 5**);
- to respond, as appropriate, to a series of item-specific probe questions requesting information on such topics as concept or question comprehension, question readability, and proxy responding; and,
- to respond to a series of general probe questions requesting information on a variety of related topics (e.g., questionnaire flow; utility of information provided in the supplement interviewer manual; unusual patterns of responding based on demographic characteristics; proxy responding).

II.B.3. *Response Distribution Analysis.* With the exception of split-panel research that compares the effects of differential question wording, it is fairly rare to find practitioners who make use of response distribution data to assess data quality—other than to note item nonresponse rates, perhaps—and few methodologists even list this analytical strategy as an evaluation method (for exceptions, see DeMaio, Mathiowetz, Rothgeb, Beach and Durant, 1993; Esposito and Rothgeb, 1997). In this research effort, cross-tabulations involving sets of supplement items were used to identify unusual/unexpected patterns of responding that could provide indirect evidence of possible measurement error or willful misreporting by respondents (e.g., highly unlikely patterns of "no" responses to key supplement items).

II.B.4. *Expectations Regarding Relationships Between Pre-survey and Post-survey Evaluation Data.* To the extent that each of the three principal evaluation methods used in this research yielded one or more quantitative indicators (i.e., the QAS *problem count*; *percentages* associated with the various interviewer and respondent behavior codes; the item-specific *rating data* provided by telephone center interviewers), it was possible to compute correlations between the various indicators to determine to what extent they were interrelated. Given the exploratory nature of this research, the only expectation held prior to analyzing these data was that positive (and possibly significant) correlations would be found between various *problem indicators* (cf. Presser and Blair, 1994; Willis,

Schechter and Whitaker, 1999). For example, considering the set of twelve supplement items, it was expected that item-specific QAS indicators (i.e., the sum of problems detected for a particular supplement item) would correlate positively with a corresponding set of values derived by summing "suboptimal" respondent behavior codes (i.e., the sum of all respondent-code percentages other than the percentage associated with the adequate-answer code). In other words, if item "x" has a high score on one problem indicator, it should also score high on another problem indicator, and vice versa.

III. Findings

The principal findings of this research, those pertaining to predictions regarding problems with specific supplement items and those pertaining to correlations between pre-survey and post-survey evaluation data, can be found in **Table 3** and **Table 6**, respectively. Other supporting data pertaining to the twelve items that comprise the cell-phone-use supplement can be found in the following tables:

- **Table 1:** Response distribution data.
- **Table 2:** QAS data.
- **Table 4:** Behavior-coding data.
- **Table 5:** Ratings of *respondent* difficulty with supplement items (Note: These ratings were *assigned by interviewers*).

III.A. Predictions Regarding Problems with Specific Supplement Items.

A total of thirty two item-specific predictions were made prior to supplement administration in February 2004 and the subsequent collection of behavior-coding and interviewer-debriefing data. The first prediction in each set of item-specific predictions was based primarily on QAS evaluation work (see subsection II.A.3 for additional information on the QAS). Other item-specific predictions in a particular set [e.g., predictions Q3 (B) and (C)] were based on information gleaned from cognitive interviews (e.g., from information summarized in various review-and-recommendation memos; see subsection II.A.1) and from SORT-team feedback (see subsection II.A.2). The latter set of predictions were far more specific than the former set, and sometimes overlapped in content [e.g., see predictions Q3 (A) and (C)]; the overlap in content, though not a serious obstacle in assessing predictive outcomes (e.g., hit; miss), should nevertheless be viewed

as a methodological flaw. Assessing the outcome of an item-specific prediction was essentially a subjective process which involved reviewing a good portion of the evaluation data available for a specific supplement item (i.e., behavior-coding and interviewer-debriefing data) and forming a judgment as to whether the available data provided sufficient evidence to confirm the prediction (see **Table 3**). There were five outcome categories:

- *Hit*: This is where a prediction regarding the existence of a problem appears to be confirmed on the basis of available evaluation data.
- *Partial Hit* (or *Partial Miss*, if you prefer): This is where a prediction regarding the existence of a problem appears to be partially confirmed on the basis of available evaluation data.
- *Miss*: This is where a prediction regarding the existence of a problem *does not* appear to be confirmed on the basis of available evaluation data.
- *Missed Problem*: This is where no prediction regarding the existence of a problem was made, but where available evaluation data suggest a problem does exist.
- *Insufficient Data*: This is where a prediction as to the existence of a problem could not be determined due to a paucity of evaluation data (e.g., low frequency of administration of a particular supplement item).

A tabulation of the outcomes associated with the thirty-two predictions is provided below. As one can see, the largest outcome category was "insufficient data" (41%), followed by "hits" (31%) and "misses" (22%). The percentage of hits, no doubt impressive if one is touting the batting average of a favorite baseball player, probably will not inspire confidence among the community of practitioners who conduct questionnaire evaluation research. Surely it should be possible to do better. We will revisit this distribution of outcomes in the discussion section.

Outcome Category	Frequency	Percentage
Hits	10	31%
Partial Hits	2	6%
Misses	7	22%
[Missed Problems]	[2]	[6%]
Insufficient Data	13	41%
Prediction Total	**32**	**100%**

III.B. Relationships Between Pre-survey and Post-survey Evaluation Data.

In order to explore relationships between pre-survey and post-survey evaluation data, an 8-by-8 correlation matrix was generated that involved the following eight quantitative indicators. [4]

- (1) QAS data (**QAS**): This indicator (the only quantitative indicator available from pre-survey evaluation work), was derived by summing the "yes" entries that correspond to the twenty-six QAS subcategories; there is one indicator associated with each of twelve supplement items (see Table 2). The higher the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (2) Interviewer-debriefing data (**HTC rating**): The average *item-specific difficulty rating* assigned by interviewers at the Hagerstown telephone center (see Table 5). The higher the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (3) Interviewer-debriefing data (**TTC rating**): The average *item-specific difficulty rating* assigned by interviewers at the Tucson telephone center (see Table 5). The higher the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (4) Interviewer-debriefing data (**Average rating**): The *average* of the group *item-specific difficulty ratings* assigned by interviewers at the both Hagerstown and the Tucson telephone centers (see Table 5). The higher the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (5) Behavior-coding data (**BC 'E'**): The *percentage* of "exact" question readings associated with the interviewer behavior codes (see Table 4). The lower the value of

[4] It is important to note, as others have (e.g., Forsyth, Rothgeb and Willis, 2004), that indicators derived from various survey evaluation methods are indirect and imperfect measures of data quality. With regard to this research, the existence of a positive relationship is *presumed* between problem indicators (like high QAS scores) and item-specific measurement error, however I do so with some trepidation. Establishing the strength of this relationship in any specific research context is an empirical issue requiring a "true-score" data source (e.g., Dykema, Lepkowski and Blixt, 1997). For a thoughtful discussion of these issues, see Willis, DeMaio and Harris-Kojetin, (1999).

this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.

- (6) Behavior-coding data (**BC 'Not-E'**): The *cumulative percentage* of all interviewer codes except the "exact" question-reading code (see Table 4). The higher the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (7) Behavior-coding data (**BC 'AA'**): The *percentage* of "adequate answer" codes associated with the respondent behavior codes (see Table 4). The lower the value of this item-specific indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.
- (8) Behavior-coding data (**BC 'Not-AA'**): The *cumulative percentage* of all respondent codes except the "adequate-answer" code (see Table 4). The higher the value of this indicator, the greater the measurement error associated with a particular questionnaire item is presumed to be.

There are twelve data points for each indicator (N=12), corresponding to the twelve supplement items. A total of twenty-eight correlations were generated, twenty of which were considered informative in the substantive sense (**Table 6**; the eight non-informative correlations have been placed in brackets).[5] Of the twenty informative correlations, four entries were significant at the .05 level or better (one-tailed test) and one entry was marginally significant (.061 level). To review correlations between pre-survey evaluation data (i.e., the QAS indicator) and the post-survey evaluation data (as decomposed into seven distinct indicators), the reader is directed to the first data column of Table 6. Although none of the corrections between the QAS and the three interviewer-rating indicators is significant, the magnitude of all three are relatively high (.375 or better) and all are positive, as expected. In contrast, though the signs of the correlations make sense (e.g., pairs of problem indicators, like the QAS and the HTC rating, correlate positively, while incongruent pairs of indicators, like the QAS and the BC 'E', correlate negatively),

[5] The two correlations involving the HTC and TTC ratings with their average score, and the two correlations between "E" and "not-E" behavior codes and "AA" and "not-AA" behavior codes are considered non-informative because the correlated sets of indicators are not independent. The four correlations involving the four behavior coding indicators are considered non-informative because they are

the magnitude of QAS correlations with the four behavior-coding indicators are relatively weak (none higher than .311). Most disappointing, given the QAS emphasis on respondent/response coding categories and subcategories (e.g., clarity; assumptions; knowledge/memory), is the weak correlation (.142) with the indicator that aggregated the percentages associated with the seven suboptimal respondent behavior codes (i.e., BC 'not-AA'). A significant correction was expected. With respect to correlations between post-survey indicators, the magnitude of the correlations between the HTC difficulty ratings (and to a lesser extent the average difficulty ratings) and the four behavior-coding indicators are all quite strong (three reaching significance) and all four have signs (positive or negative) in the expected direction. In contrast, three of four correlations involving the TTC difficulty ratings and the behavior-coding indicators are relatively weak, only the correlation with the "exact reading" behavior code approaches respectability (-.357).

As noted (see footnote 5), the four correlations involving behavior coding indicators are considered non-informative because these indicators represent aggregate measures. To remedy this situation, all item-specific behavior coding data were recoded so that relationships between question-asking and question-answering behavior could be assessed (cf. Dykema, Lepkowski and Blixt, 1991). One the interviewer side, an exact question reading ("E" code) was recoded as "0" (an "optimal" reading) and any other interviewer behavior was recoded as "1" (a "suboptimal" reading). On the respondent side, an adequate answer ("AA" code) was recoded as "0" (an "optimal" response) and all other respondent codes were recoded as "1" (a "suboptimal" response)—let's call this *Condition One*. [A second set of correlations was also computed that relaxed the criteria for an optimal question reading. In this condition, an exact question reading ("E" code) and a minor change in wording ("mC" code) were both recoded as "0" and any other interviewer behavior was recoded as "1". Let's call this *Condition Two*.] Correlations between binary interviewer and respondent codes were then computed for all twelve items individually (**Table 7**) and also for the full set of twelve items as a whole. As can be seen from Table 7, correlations between interviewer and respondent behaviors varied

---

aggregate indicators (i.e., gross percentages) and, as a result, are not capable of shedding light on the nature of the relationship between question-asking and question-answering behavior across individual cases.

widely, and in several cases the correlations were actually *negative*. For example, with respect to supplement item Q2 (correlation: -.219; N=58 cases), even though an optimal question reading was followed by an optimal response in the majority of cases (36), there were 15 cases where an optimal reading was followed by a suboptimal response and 7 cases where a suboptimal reading was followed by an optimal response; for this item, there was not a single case where a suboptimal reading was followed by a suboptimal response. When the recoded data for all twelve items are combined and correlations between binary interviewer and respondent codes are computed, the results under both testing conditions suggest that there is no linear relationship between interviewer and respondent behaviors: For Condition One, the correlation is .015 (p=.796) and for Condition Two, .048 (p=.419). These weak correlations between interviewer and respondent codes were unexpected and seem counterintuitive. When interviewers substantially alter the wording (and, in some cases, the meaning) of a well-designed survey question, most survey practitioners presumably would anticipate some sign of *turbulence* on the respondent side of the interaction (e.g., inadequate answers; requests for clarification; explanations of some sort)—perhaps not in every case, but in general. And the opposite might also be expected: When the interviewer reads the question as worded, few signs of turbulence should be noted. In this study, as the correlations above suggest (and as cross-tabulations confirm), neither expectation was fully realized. When interviewers deviated from question scripts, there were signs of turbulence on the respondent side about 25 to 30% of the time. When interviewers read questions as worded, there were signs of turbulence on the respondent side about 23% of the time.

IV. Discussion

How one feels about the value of this research effort and its findings will probably depend on how one views the questionnaire-design-and-evaluation process more generally. Those who view the process as eight parts “science” and two parts “art” will probably be disappointed with this research effort/approach. Those who view the process as four parts “science”, four parts “art” and two parts “negotiation” (which is how I tend to view the process, at present) will probably have a somewhat different reaction to this work. For those of either mindset who feel there is something worth pursuing in the

somewhat ambiguous findings reported herein, let me try to situate this research effort and its findings within a broader context.

IV.A. An Organizational Framework and More on the Evaluation Methods.

In a series of written works (e.g., Esposito, 2003, 2004a, 2004b), I have proposed and elaborated upon an organizational framework that relates various phases of the questionnaire-design-and-evaluation process to various sources of measurement error (see Attachments, **Table A-3**). This pragmatic framework may prove useful, at least in part, in our attempts to understand when and why predictions regarding item-specific problems prove accurate or inaccurate, and when and why correlations between various summary indicators are strong, while others are weak.

The *design-and-evaluation process* is thought to comprise eight phases: Four *core phases* (P1: observation; P3: conceptualization; P5: operationalization; and P7: administration) and four corresponding *evaluation phases* (P2, P4, P6 and P8, respectively). This research effort focuses on the latter three phases of the process, P6 through P8. With respect to *sources of measurement error*, five are specified (Esposito, 2004a; cf. Groves, 1989): (1) questionnaire content specialists; (2) questionnaire design specialists; (3) interviewers; (4) respondents; and (5) mode. Each *class of evaluation methods* (e.g., interaction analysis; interviewer debriefing), and each specific evaluation *technique* (e.g., behavior coding; focus group, respectively), can be thought to comprise a *core set of components/elements* and a corresponding *context-specific set of instantiations*. [6] The former refers to a method-specific set of procedural tasks—what gets done, when, and how, *in most cases*; the latter refers to those contextual features associated with the implementation of a particular method—the who, where, why and how, *in specific cases*—that distinguish that application from all others. Whenever a particular method is employed in evaluation research, it is important to note which components are

[6] For example, behavior coding can be thought to comprise a set of six components: (1) natural survey context, (2) observation of interviewer and/or respondent behavior, (3) survey questions/questionnaire; (4) coding framework; (5) coders; and (6) data analysis (see Esposito, Rothgeb and Campanelli, 1994). Drawing on variants of schema theory (e.g., Schank and Abelson, 1977), evaluation methods could be (and probably *should be*) viewed as *scripts* that vary within and between classes of methods with respect to how standardized/formalized each has become.

implemented (if the method has become standardized, this tends to be fairly consistent across practitioners and organizations) and also carefully observe how each element is instantiated (which essentially relates to the manner and degree to which the five sources of measurement error noted above are involved in the process of implementing the method).[7] For each method used (and compared), one would ideally want to consider the following types of questions:

- (1) To what extent, and in what manner (e.g., direct vs. indirect participation), have *questionnaire content specialists* been involved in implementing a particular method? Have key survey concepts been clearly defined? Have item objectives been clearly specified? Have conceptual definitions and item objectives been made available to the individuals/groups who have been asked to generate evaluation data/information?

- (2) To what extent, and in what capacity (e.g., design, evaluation, or both), have *questionnaire design-and-evaluation specialists* been involved in implementing a particular method? What knowledge do these specialists possess with respect to key survey concepts and item objectives? What level of expertise do they possess with respect to designing and evaluating survey questionnaires?

- (3) To what extent, and in what capacity (e.g., survey administration vs. research collaborators/participants), are interviewers involved in implementing a particular method? How representative are these interviewers of the population of interviewers who will be conducting the target survey?

- (4) To what extent, and in what capacity (e.g., survey participants vs. research collaborators), are respondents involved in implementing a particular method? How representative are these respondents of the population of individuals who are sampled for participation in the target survey? Does the evaluation method focus on internal mental processes (e.g., cognition; emotion; motivation), external observable/measurable/codable behaviors (e.g., response latency; manifest signs of uncertainty or confusion), or a combination of both?

---

[7] One could further characterize various evaluation methods by making use of Forsyth and Lessler's (1991) thought-provoking taxonomy of cognitive laboratory methods. Their framework distinguishes evaluation methods according to two dimensions: *Task timing* (concurrent, immediate, delayed and unrelated) and *attention control* (unrestricted, directed, external and dynamic).

- (5) To what extent does the mode in which the evaluation method is implemented simulate the conditions in which the actual survey will be administered? In what situations would it be accurate to classify an evaluation method as "modeless," that is, one that does not involve actual interactions between interviewers and respondents in a real or simulated interview context?

Before moving on to a discussion of findings, let's consider some of the more salient aspects of our pre-survey and post-survey evaluation methods.

IV.A.1. *Cognitive interviews*. As noted in subsection II.A.1, of the three pre-survey evaluation methods, only the cognitive interviews, which were conducted over the telephone for the most part, actually involved an "interviewer" (a behavioral scientist) asking a "respondent" (the research participant) draft supplement questions; and, even then, the interview experience was far different from an actual CPS interview (e.g., most research participants are paid, and sets of scripted and unscripted probes repeatedly interrupt the normal question-asking sequence). And, of course, one of the defining characteristics of cognitive interviews is that they are designed to gather information on a subset of human mental processes—cognitive processes (e.g., comprehension and retrieval) as opposed to motivational or emotional processes—that individuals employ in answering survey questions; however, the relatively small samples of individuals who volunteer to participate in cognitive interviews generally can not be considered representative of the sample of individuals who are selected each month to participate in national surveys, like the CPS.

IV.A.2. *SORT-Team Review.* As noted in subsection II.A.2, this evaluation method, which involved a group of experienced CPS interviewers, is probably best classified as a form of (informal) expert review. Team members were asked to provide comments on a near-final draft of the cell-phone-use questionnaire; a set of stimulus questions were provided, but no formal coding categories. Insofar as the supplement instructional memorandum for CPS interviewers was not finalized until some time afterwards, it is unlikely that team members had access to information regarding key survey concepts or item objectives. So, relative to the other methods used in this research effort (e.g., the cognitive interviews; behavior coding), this review process should be considered a fairly

subjective process (i.e., one based on a particular team member's experience with similar questions/questionnaires). [8]

IV.A.3. *QAS.* As noted in subsection II.A.3., the present author was responsible for using the QAS method to evaluate the twelve items on the supplement questionnaire; however, because I had listened to audiotapes of the cognitive interviews and had reviewed SORT-team feedback several months prior to completing the QAS task, the implementation of this method cannot be considered completely independent of information obtained from the other two methods. For these reasons, it is somewhat difficult to classify the QAS evaluation process on the subjectivity-objectivity dimension—close to the middle, perhaps, but more to the subjectivity side. One of the obvious strengths of the QAS is that it has explicit rating categories and subcategories and well-written documentation (e.g., instructions and examples) on how to assign yes-no codes. While the QAS is designed to help practitioners identify problematic questions, a careful review of its categories and subcategories would seem to suggest a greater emphasis on identifying problems that would affect respondents (e.g., instructions; clarity; assumptions; knowledge/memory; sensitivity), as opposed to those that would affect interviewers (e.g., reading). While anyone could probably learn to apply the QAS fairly reliably, skilled survey practitioners (e.g., cognitive psychologists) probably would have a substantial advantage both in reliability and accuracy.

IV.A.4. *Behavior Coding.* As noted in subsection II.B.1, the present author was responsible for coding interactions between interviewers and respondents during administration of the cell-phone-use supplement in February 2004; a total of sixty cases were coded. Of the various evaluation methods used in this research effort, behavior coding, as a process, probably should be considered as the most objective/empirical method. The coder had carefully reviewed the supplement instructional memorandum (i.e., possessed sufficient although not extensive knowledge of concepts and objectives), had considerable experience with behavior codes and procedures, and had listened to sixty supplement interviews without influencing or altering the behavior of survey

[8] Making use of a subjectivity-objectivity dimension to characterize and compare survey evaluation methods is not original to this paper. Willis, Schechter and Whitaker (1999, p. 32) refer to the "continuum of objectivity" in their work and use it as a basis for formulating hypotheses regarding the magnitude of correlations between various evaluation methods.

participants. On the downside, coding was done while interviews were in progress, which generally results in some interactions being miscoded or missed completely, and was only conducted at telephone centers, which may not be representative of field-based telephone and personal interviews, the latter accounting for 85-to-90% of all CPS interviews.

IV.A.5. *Interviewer Debriefing.* As noted in subsection II.B.2, the present author was responsible for debriefing CPS interviewers at two telephone centers; as part of that process, interviewers completed a standardized rating task focusing on items they had identified as problematic at the outset of the debriefing session. Given the qualitative information provided by interviewers during the debriefing sessions (which essentially documented their observations and opinions of how well the supplement questions had worked during actual interviews) and the quantitative data that was generated by the rating task (which involved estimating how difficult it was for *respondents* to provide adequate answers to problematic supplement items), the interviewer-debriefing process probably should be placed closer to the subjectivity side of the subjectivity-objectivity dimension. Interviewers had been given instructional materials for the supplement (concepts and objectives) and, as far as can be determined, were skilled/experienced at doing their jobs. And while debriefing interviewers has proved very useful in documenting *manifest* problems with survey questions, an interviewer's capacity to detect problems with underlying mental processes is limited, relatively speaking.

Let's revisit the findings reported earlier and, after noting some obvious methodological shortcomings with the prediction-formulation-and-confirmation process, try to determine to what extent the framework might be useful as a device for explaining, at least in part, what was observed.

IV.B. Predictions Regarding Problems with Specific Supplement Items.

As noted in subsection III.A, only about 31% of the predictions (10 of 32) regarding "problems" with specific supplement items appear to have been confirmed; 22% of the predictions (7 of 32) were "misses" and 41% (13 of 32) could not be confirmed or disconfirmed due to insufficient data. Could we have done better? Absolutely.

Methodologically, predictions could have been formulated more discretely/precisely and unambiguous criteria for the various outcome categories (e.g., hits; misses) could have been specified ahead of time. Moreover, outcome data could have been corroborated by a research associate. More pragmatically, I probably should have known/anticipated that there would be a deficit of evaluation data/information available for making predictions about infrequently asked supplement items (i.e., VER1; VER2; Q1a; Q1b; Q2e). These shortcomings notwithstanding, is there anything more that can said with regard to the seven predictions classified as "misses" [i.e., Q1a (A) and (B); Q1b (A); Q2 (B); Q2a (A) and (B); and Q2c (A)]? How might we explain these missed predictions? What insights, if any, might we gain by referring to interrelationships specified within the framework?

Though not to be considered mutually exclusive, the following set of explanations would appear relevant both to pre-survey and post-survey evaluation work.

- *Ecological realism/validity.* Methods that simulate or capture the "real world" of surveys (i.e., actual interactions between interviewers and respondents in natural survey contexts) may be more efficient than "technical" methods (e.g., those based on a comprehensive set of design issues and/or cognitive categories) at diagnosing and confirming *manifest problems* with survey questions/questionnaires. However, more technical methods—like the QAS—may be more efficient at identifying a broad range of potentially important *latent problems* (e.g., faulty assumptions; reference period issues; recall or retrieval problems) and, as a result, will require more sensitive methods to confirm such problems (e.g., post-administration, response-specific debriefing questions). In this research, predictions based largely on QAS data accounted for three "hits", one "partial hit", and four "misses." Why the mixed results? It is possible that the QAS, when used in pre-survey evaluation work, identified a relatively large number of problems (both of the latent and manifest variety) that were not confirmed in post-survey evaluation work, because: (a) post-survey evaluation methods (behavior coding and interviewer debriefing) gather data/information from actual survey contexts and, as a result, are more likely to detect manifest problems and miss latent problems; (b) the QAS may require a more sensitive post-survey evaluation method—like response-specific respondent

debriefing questions—for detecting and/or confirming the existence of latent problems; and/or (c) the method may be insensitive to the actual number of survey participants who are likely to be affected by the problems identified—in other words, the criteria for assigning a "yes" entry to various problem subcategories may not reflect the level of difficulty actually experienced by most individuals.

- *The number and mix of methods.* It is taken as axiomatic by some survey methodologists (e.g., Oksenberg, Cannell and Kalton, 1991; Esposito and Rothgeb, 1997, pp. 562-566) that a strategic mix of evaluation methods—that is, a mix that allows the practitioner to observe and assess the influence of multiple sources of measurement error (i.e., content and design specialists; interviewers and respondents; mode)—is superior to: (1) a single-method evaluation strategy for diagnosing problems; or (2) an unbalanced mix of methods that excludes certain sources of error or a mix that is highly redundant. For example, in this research, the mix of methods used in pre-survey evaluation work drew heavily on the appraisals and reviews of "experts" (i.e., QAS and SORT comments, respectively). In contrast, there was relatively little data/information available on what was likely to happen in the context of an actual supplement interview. A more balanced mix of pre-survey evaluation methods may have resulted in a different set of predictions, and possibly fewer predictive "misses." One also has to consider what changes to a draft questionnaire have been made (and what changes have *not* been made) on the basis of evaluation findings. To the extend that a particular method's findings have led to constructive changes, that method loses some of its potential for predicting the types of problems that are apt to arise during survey administration.
- *Satisficing and Misreporting.* Respondents are not always motivated to exert the effort necessary to perform the various tasks that would be required to answer survey questions accurately/optimally; this particular form of survey responding, the general prevalence of which is not known, has come to be called *satisficing* (Krosnick, 1991). Though we tend to associate satisficing with respondents, a more thorough analysis would probably show that there are multiple causal antecedents (e.g., a rapid pace of interviewing; poorly conceived/designed questionnaire items; an irrelevant or uninteresting survey topic), all of which presumably interact, to explain such

behavior. That said, some respondents, motivated to end the interview as quickly as possible, will offer little resistance to answering questionnaire items that are ambiguous or appear to be problematic in some other sense (e.g., heavy response burden; difficult mapping to respondent's situation). When this happens in the course of post-survey evaluation work, item-specific problems are apt to be missed or undercounted, and predictions made on the basis of pre-survey evaluation work (e.g., cognitive interviews; expert appraisals) are apt to be compromised. For example, supplement item Q1b asks about taking incoming calls on a landline number, but was mistakenly asked by some telephone-center interviewers in February 2004 because they had incorrectly entered the wrong precode to a prior check-item. As it happened then, a relatively small percentage of respondents were asked if they took incoming calls on a landline number, *after they had just answered an incoming call on their landline number and were speaking to the interviewer on that landline number*. Eight of these cases were monitored during behavior coding and not a single respondent commented on the absurdity of the question. Interviewers recognized the problem, but respondents were mute on the issue. There are several other items that, for various reasons (e.g., unspecified reference period; ambiguous intent or wording; unspecified assumptions regarding cell or telephone use), were identified as problematic prior to supplement administration, and these items precipitated little or no reactions from respondents. One can only speculate on the quality of data being collected in such circumstances. *Misreporting*, the conscious reporting of inaccurate information, is also a potentially serious problem, and the magnitude of the problem is often difficult to estimate. One indirect means of doing do is by reviewing response-distribution data and analyzing cross-tabulation data to uncover highly unlikely response patterns (see subsection II.B.3). For example, 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? Well, I suppose there

are plausible reasons (e.g., no friends or family; only communicate via computer); however, this group of respondents may also wish to avoid being contacted by individuals who conduct surveys—and if so, they may misreport. But 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 information provided by such a question, not to mention the response latency, has the potential to be very useful.

- *Questionnaire items with low frequencies*. When survey questionnaires involve branching, as many governmental and private-sector surveys do, some items will be asked less frequently than others. When this happens, and when there are constraints on how much and what types of evaluation data can be collected, the likelihood increases that predictions regarding potential problems with low-frequency items may not be confirmed (i.e., as a consequence of the low "effective sample size"; see Willis, DeMaio and Harris-Kojetin, 1999, pp. 145-146); a number of problems may be observed and reported, but not enough to cross the threshold for classifying such items as problematic. With regard to this research, three of the seven predictions that were classified as "missed" involved low-frequency items [i.e., Q1a (A) and (B); Q1b (A)]; and, as previously noted, 41% (13 of 32) of the predictions could not be assessed due to insufficient data.

IV.C. Relationships Between Pre-survey and Post-survey Evaluation Data.

As the reader may have surmised, most of the explanations discussed above regarding predictive outcomes can be applied or modified for use in discussing correlations *between* pre-survey and post-survey evaluation data (i.e., quantitative indicators); the one exception is the *number/mix-of-methods*, because these correlation data involve one-to-one comparisons, not predictions based on multiple methods.

- *Ecological realism/validity.* Even though the QAS has been characterized as a "technical" evaluation method (as opposed to methods that are implemented within

natural or simulated survey contexts), it does have a significant orientation towards issues that affect respondents and that emphasis may account for the fairly strong correlations with the three interviewer-rating indicators (see Table 6, column one correlations). However, that technical character does not bode well for its relationships with the four behavior-coding indicators, not even with the "BC not-AA" indicator (r = .142), which represents an aggregate of suboptimal *respondent* behavior codes. The QAS propensity to identify one or more problems with almost any questionnaire item (see Rothgeb, Willis and Forsyth, 2001) and its predisposition towards detecting problems associated with unobservable mental processes, may help to explain why correlations with behavior-coding data, and its predisposition towards manifest problems of the interactive sort, are not higher.

- *Satisficing and Misreporting.* Although neither of these behavioral strategies are relevant with respect to pre-survey evaluation data—the QAS was implemented with enthusiasm by the present author, whose most serious shortcoming in using this particular method may have been inexperience—satisficing and misreporting may have been an indirect factor affecting relationships involving the post-survey evaluation data if these behaviors actually dampened the likelihood and frequency of problematic exchanges between interviewers and respondents during supplement administration in February 2004. There is some evidence of satisficing, in my opinion, with respect to supplement item Q3, which suffered from ambiguous wording and underdeveloped item specifications. Q3 asks about cell-phone use, relative to landline telephone use, and was essentially impossible to answer in some situations (e.g., large families with multiple cell-phone users). Of 14,451 responses to this question, only 211 were coded as "don't know" and only 36 were coded as refusals—a combined percentage of only 1.7%. The demands on working memory could be overwhelming (once it had been determined how much arithmetic the question actually requires) and it seems unlikely that respondents could have satisfied those demands in the very brief amount of time it took for most of them to provide an answer.
- *Questionnaire items with low frequencies (and analyses with low statistical power).* Especially with respect to behavior-coding data, when a particular questionnaire item

is asked infrequently, the data for interviewer and respondent exchanges tends to be unreliable/unstable, relative to items that are asked more frequently; and a lot depends on the circumstances of the few cases that are coded (e.g., interviewer skill; the respondent's knowledge and circumstances). Evaluation data (i.e., ratings) provided by interviewers can be considered somewhat more reliable, because their judgments are spread over a larger number of cases (e.g., monthly interview caseloads may involve up to 40-to-50 households). Correlations between the QAS and the two groupings of indicators for behavior coding (which range from a low of .142 to a high of .311) and interviewer ratings (low of .375, high of .471) appear consistent with these generalizations (see Table 6). Lastly, the fact that there were only twelve supplement items, and thus only twelve data points for computing correlations, means that the statistical power associated with each of our correlations was very low. Especially problematic for low-frequency items is the adverse effect that a single outlier can have on correlation magnitudes. The combination of low- frequency items (unreliable/unstable data) and low power is particularly lethal when conducting correlation analyses.

IV.D. Relationships Between Problem Indicators: Post-survey Evaluation Data.

Let's now briefly consider relationships between problem indicators derived from post-survey evaluation data only.

- *Ecological realism/validity.* Though interviewer rating data and behavior-coding data are both rooted in interviewer-respondent exchanges in natural survey contexts, there are some differences between the two sets of quantitative indicators that may help to explain differences in relationship/correlation magnitudes. For example, the ratings data provided by interviewers pertain to item-specific difficulty levels that they *attributed* to the respondents they had interviewed (a relatively subjective process), whereas behavior-coding data are based on the empirical observations of interviewer-respondent exchanges by an independent monitor/coder (a relatively objective process). Also, with respect to the characteristics of the interviewers who provided the ratings, the two groups of interviewers differed in three respects: (1) ethnic background (i.e., by request, five/half of the interviewers from Tucson were Hispanic; none from the Hagerstown group were), (2) total experience as interviewers (i.e.,

averages: Tucson group, 5.44 years; Hagerstown group, 6.75 years); and (3) gender composition (i.e., Tucson group, six women, four men; Hagerstown group, ten women). These differences notwithstanding, one might expect fairly high correlations between ratings data and behavior-coding data, especially with respect to the respondent behavior codes. The results were mixed, however. Unexpectedly, there were dramatic differences between the two groups of interviewers with respect to correlations between the ratings scores and the four behavior-coding indicators: Three of four correlations are significant for the Hagerstown group, none for the Tucson group. These data suggest that Hagerstown interviewers were more sensitive to the problems being experienced by respondents—this in spite of the imperfections associated with both evaluation methods. Experience may have been a key factor, and other interviewer feedback suggests that Hispanics may have had less exposure to the supplement, due to the oversampling of Hispanic households for another CPS supplement that was being conducted concurrently with the cell-phone-use supplement. It is not known what effect gender differences between the two groups may have affected these correlation values. In a review of the focus group literature, Bischoping and Dykema (1999) look at this technique from a social psychological perspective and cite studies which suggest that "… women's input in focus groups would indeed be enhanced by participating in all-female groups (p. 499)." A similar finding would appear to hold for minority participants. Though enlightening, these general findings would not be very helpful in explaining why correlations between interviewer ratings and behavior coding data differed for the two groups of interviewers (Hagerstown versus Tucson), because both groups identified about the same number of supplement items as problematic and rating data were collected *before* interviewers were provided with the opportunity for a full discussion of problematic items. A more likely gender-based explanation for these differences would be gender effects related to listening behavior and empathy. Unexpected, too, were the very low correlations between case-specific interviewer and respondent behavior codes (see Table 7). In spite of some reservations about specific supplement items, I had anticipated fairly strong positive correlations between interviewer and respondent behaviors, assuming that optimal interviewer behavior would facilitate

optimal respondent behavior and that suboptimal interviewer behavior would precipitate suboptimal respondent behavior. These strong correlations did not materialize and the reason is not immediately apparent. One unexplored possibility is that interviews conducted from telephone/CATI centers, which are regularly monitored by supervisors and draw on a more receptive group of respondents, are qualitatively different from interviews conducted in the field. Other possibilities are noted below.

- *Satisficing and Misreporting.* Neither of these behaviors would appear to account for the general *pattern* of correlations between problem indicators, though they might have played a role more specifically in the low correlations that were observed between interviewer and respondent behavior codes. With regard to the latter, the reality may be that there are no strong positive correlations to be found between interviewer and respondent behavior codes, especially if items are poorly designed and/or if satisficing suppresses the reactions of respondents to suboptimal items. Consider two additional explanations for why strong positive correlations may be rare. First, the communication of intended meaning requires more than survey questions being read exactly as worded. As Suchman and Jordan (1990) have noted: "Stability of meaning, …, requires the full resources of conversational interaction (p. 233)." In other words, an interviewer behavior that is coded as a "major change" in question wording by an independent observer may not necessarily constitute an actual change in communicated meaning—in theory, the full interactional context would need to be considered before such a judgment could be made. And second, an "adequate answer" to a given survey question does not necessarily constitute as an *accurate* answer to that question; validation data or post-administration probing may indicate that the response is actually inaccurate—and that inaccuracy may not be entirely attributable to the respondent (e.g., ambiguous question wording; inadequate item specifications; insufficient processing time due to rapid interviewing pace). For example, Dykema, Lepkowski and Blixt (1997) conducted an illuminating validation study in which they investigated relationships between interviewer and respondent behavior codes, on the one hand, and response accuracy, on the other; data were analyzed using three logistic regression models. On the basis of findings from their

second model, in which they aggregated "suboptimal" [my adjective, here and below] interviewer codes and "suboptimal" respondent codes as separate variables prior to analysis, the authors conclude: "Thus, none of the errors made by interviewers appears to be systematically related to accuracy in our analysis. However, in eight of the ten tests shown in [Table 12.3, Panel A], respondent codes [i.e., a summation of codes for interruptions, uncertainty, qualified answer, uncodeable response, don't know, refusal] are positively associated with inaccurate responses (p. 301)."

- *Questionnaire items with low frequencies (and analyses with low statistical power).* Low frequencies (for a number of supplement items) and low statistical power (N=12 items) would appear to be the most plausible explanation for the low correlations between the interviewer behavior codes and the respondent behavior codes.

IV.D. Closing Remarks

Sometimes, not finding what you expect to find—be they expectations regarding item-specific problems during survey administration or relationships between method-based problem indicators—can be a viewed as a positive outcome, if we learn something useful along the way. That "something" can be a more efficient way to implement an evaluation method, or an insight sparked by an unconfirmed expectation, or an enhanced appreciation of the complexity of the question-and-answer process. "Success" in this particular domain (and in any research domain involving human behavior) is elusive and sometimes misleading. The ultimate goal is understanding the process, and sometimes we can move a step closer to that goal by seizing opportunities for research that are something less than optimal.

References

Bischoping, K., and Dykema, J. (1999). "Towards a Social Psychological Programme for Improving Focus Group Methods of Developing Questionnaires." *Journal of Official Statistics*, 15: 495-516.

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.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Surveys*. New York: Wiley, 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.

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

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." 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, 287-310.

Esposito, J.L. (2004a). "Iterative, Multiple-Method Questionnaire Evaluation Research." *Journal of Official Statistics*, 20: 143-184.

Esposito, J.L. (2004b). "With Regard to the Design of Major Statistical Surveys: Are We Waiting Too Long to Evaluate Substantive Questionnaire Content?" *QUEST2003: Proceedings of the Fourth Conference on Questionnaire Evaluation Standards*. Mannheim: ZUMA, 161-171.

Esposito, J.L. (2003). "A Framework for Relating Questionnaire Design and Evaluation Processes to Sources of Measurement Error." *Statistical Policy Working Paper 37*. Federal Committee on Statistical Methodology, Washington, DC: 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, 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.

Forsyth, B., Rothgeb, J.M. and Willis, G. (2004). "Does Pretesting Make a Difference?" In S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Surveys*. New York: Wiley, 525-546.

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

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

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

Gerber, E.R., and Wellens, T.R. (1997). "Perspectives on Pretesting: 'Cognition' in the Cognitive Interview." *Bulletin de Methodologie Sociologique*, 55, 18-39.

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

Hess, J. Singer, E. and Bushery, J. (1999). "Predicting Test-Retest Reliability from Behavior Coding." *International Journal of Public Opinion Research*, 11: 346-360.

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

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

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

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

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

Schank, R.C., and Abelson, R.P. (1977). *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

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

Willis, G.B. (2004). "Cognitive Interviewing Revisited: A Useful Technique, in Theory?" In S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Surveys*, New York: Wiley, 23-43.

Willis, G., DeMaio, T., and Harris-Kojetin (1999). "Is the Bandwagon Heading to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques." In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds.), *Cognition and Survey Research*, New York: Wiley, 133-153.

Willis, G., Schechter, S., and Whitaker, K. (1999). "A Comparison of Cognitive Interviewing, Expert Review and Behavior Coding: What Do They Tell Us?" *Proceedings of the ASA Section on Survey Research Methods.* Alexandria, VA: American Statistical Association.

Willis, G.B. and Lessler, J.T. (1999). "Question Appraisal System, QAS-99." Instructional Manual. Rockville, MD: Research Triangle Institute.

TABLES

Table 1. Cell Phone Use Supplement, February 2004: Supplement Items and Response Distribution Data [Edited but Unweighted Data]

	Supplement Questions
N=30,523	**Q1:** 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. How many different landline telephone numbers does your household have?
Percent	
5.0	<0> Zero [1539 cases]
83.5	<1> One
9.7	<2> Two
1.4	<3> Three
0.31	<4> Four
0.06	<5> Five
0.02	<6> Six
0.01	<7> Seven
	[Don't know=10 cases]
	[Refused=35 cases]
N=1553	**VER1:** 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?
Percent	
100.0	<1> Yes
*	<2> No [* 'No' responses were recycled back to Q1. The frequency of 'no' response is unknown.]
N=3514	**VER2:** I just want to verify that your household has [fill Q1] distinct telephone NUMBERS: Is that correct?
Percent	
100.0	<1> Yes
*	<2> No [* 'No' responses were recycled back to Q1. The frequency of 'no' response is unknown.]
N=3481	**Q1a:** Excluding any numbers used only for faxes and computers, how many of these [fill Q1] landline telephone numbers are used for incoming calls?
Percent	
0.43	<0> Zero [15 cases]
48.2	<1> One
46.0	<2> Two
4.5	<3> Three
0.66	<4> Four
0.14	<5> Five
0.14	<6> Six
	[Don't know=6 cases]
	[Refused=3 cases]

N=6031 **Q1b:** 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?

Percent	
83.6	<1> Yes
16.4	<2> No [989 cases]
	[Don't know=8 cases]
	[Refused=24 cases]

N=30,184 **Q2:** [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?

Percent	
62.7	<1> Yes
37.4	<2> No
	[Don't know=87 cases]
	[Refused=256 cases]

N=18,830 **Q2a:** [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 >1]

Percent	
0.30	<0> Zero [57 cases]
51.4	<1> One
36.5	<2> Two
8.6	<3> Three
2.6	<4> Four
0.46	<5> Five
0.07	<6> Six
0.02	<7> Seven
---	<8> Eight
0.01	<9> Nine
	[Don't know=21 cases]
	[Refused=53 cases]

N=9089 **Q2b:** 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?

Percent	
4.7	<0> Zero [423 cases]
9.6	<1> One
65.8	<2> Two
14.6	<3> Three
4.5	<4> Four
0.76	<5> Five
0.09	<6> Six
0.02	<7> Seven
---	<8> Eight
0.01	<9> Nine
	[Don't know=8 cases]
	[Refused=2 cases]

N=8509 **Q2c:** How many of the [fill Q2a] cell phone numbers are answered by more than one household member?

Percent	
46.9	<0> Zero
19.0	<1> One
28.4	<2> Two
4.4	<3> Three
1.3	<4> Four
0.11	<5> Five
0.01	<6> Six
0.02	<7> Seven

[Don't know=10 cases]
[Refused=5 cases]

N=9649 **Q2d:** Do you [fill (or members of your household) if NUMHOU > 1] regularly answer this cell phone number?

Percent	
73.0	<1> Yes
27.0	<2> No

[Don't know=12 cases]
[Refused=12 cases]

N=6367 **Q2e:** Is this cell phone number answered by more than one household member?

Percent	
38.1	<1> Yes
61.9	<2> No

[Don't know=5 cases]
[Refused=4 cases]

N=14,204 **Q3:** 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 …

Percent	
8.0	<1> All or almost all calls
24.0	<2> More than half,
35.8	<3> Less than half, or
32.2	<4> Very few or none?

[Don't know=211 cases]
[Refused=36 cases]

Table 2. Questionnaire Appraisal for Cell Phone Use Supplement [QAS-99; Willis and Lessler, 1999]

		Q1	VER1	VER2	Q1a	Q1b	Q2	Q2a	Q2b	Q2c	Q2d	Q2e	Q3
1	**Reading**												
1A	What to Read	N	N	N	N	N	N	N	N	N	N	N	N
1B	Missing Information	N	N	N	N	N	N	Y	Y	N	N	N	N
1C	How to Read	N	N	N	N	N	N	N	N	N	N	N	N
2	**Instructions**												
2A	Conflicting/Inaccurate	Y	N	N	N	N	Y	Y	N	N	N	N	N
2B	Complicated	Y	N	N	Y	Y	Y	Y	N	N	N	N	N
3	**Clarity**												
3A	Wording	Y	N	N	Y	Y	Y	Y	Y	N	N	N	Y
3B	Technical Terms	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3C	Vague	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3D	Reference Periods	N	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y
4	**Assumptions**												
4A	Inappropriate	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y	N
4B	Constant Behavior	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
4C	Double Barreled	N	N	N	N	Y	Y	N	Y	N	Y	N	Y
5	**Knowledge/Memory**												
5A	Knowledge	N	N	N	Y	Y	Y	Y	Y	Y	Y	N	Y
5B	Attitude	N	N	N	N	N	N	N	N	N	N	N	N
5C	Recall	N	N	N	N	N	N	N	N	N	N	N	Y
5D	Computation	N	N	N	N	N	N	N	N	N	N	N	Y
6	**Sensitivity**												
6A	Sensitive Content	Y	Y	N	N	N	N	N	N	Y	N	Y	N
6B	Sensitive Wording	N	N	N	N	N	N	N	N	N	N	N	N
6C	Socially Acceptable	N	N	N	N	N	N	N	N	N	N	N	N

[continued]

		Q1	VER1	VER2	Q1a	Q1b	Q2	Q2a	Q2b	Q2c	Q2d	Q2e	Q3
7	**Response Categories**												
7A	Open-Ended Question	N	N	N	N	N	N	N	N	N	N	N	N
7B	Mismatch	N	N	N	N	N	N	Y	N	N	N	N	N
7C	Technical Term(s)	N	N	N	N	N	N	N	N	N	N	N	N
7D	Vague	N	N	N	N	N	N	N	N	N	N	N	Y
7E	Overlapping	N	N	N	N	N	N	N	N	N	N	N	Y
7F	Missing	N	N	N	N	Y	N	N	N	N	N	Y	Y
7G	Illogical order	N	N	N	N	N	N	N	N	N	N	N	N
8	**Other Problems**												
	Total "Yes"	**7**	**3**	**2**	**8**	**10**	**9**	**11**	**9**	**7**	**7**	**7**	**12**

Table 3. Cell-Phone-Use Supplement Items: Predictions, Data and Assessments

Q1. 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.

How many different landline telephone numbers does your household have?

[Skip Instructions: If number equals 0, Go to VER1. If number equals 1 and telephone interview, Go to Q2. [If number equals 1 and personal interview, Go to Q1b. If number greater than 1, Go to VER2.]

Prediction Q1 (A): Problems anticipated with instructions (e.g., confusion with respect to reporting telephone numbers vs. the number of telephones), clarity (technical terms, such as "landline" and "fax machine"; "household," especially from different ethnic perspectives), and assumptions.

Prediction Q1 (B): Older, more isolated respondents will have difficulty with this question (and subsequent questions, as well) due to age-related cognitive, sensory and social deficits—and due to a general lack of experience with recent communication technology (e.g., cell phones, fax machines, personal computers).

Evaluation Data/Information

Behavior Coding: Interviewers experienced a fair amount of difficulty reading Q1 as worded (exact code: 62%); in about four of ten cases, they read this question with minor changes (17%) or major changes in wording (22%). Most of the major changes involved the interviewer not reading definitional material. Respondents provided adequate (though not necessarily accurate) answers 95% of the time; however, about one-fifth of the time they interrupted the interviewer while the question was being read (interruption code: 7%) or felt the need to elaborate on the answer provided ("other" code: 15%). In one case, the respondent asked what the term "landline" meant and the interviewer answered by saying: "A phone plugged into the wall." The respondent then started counting out loud: "1, 2, 3 …" Hearing this, the interviewer asked if all the phones had one number [which is an ambiguous probe] to which the respondent said "yes". Ian answer of "one" was probably coded, but it is not clear if that was an accurate answer. In another case, a respondent interrupted and asked why the Census Bureau was asking about telephones; the interviewer struggled to provide an explanation (though an explanation was provided in their instructional materials).

Interviewer Debriefing: Of the twelve supplement items rated, Q1 placed fifth with respect to its average difficulty rating (1.78; see **Table 5** for rating task and scale). Nine (of twenty possible) interviewers recorded twenty-one comments on their log forms. Several interviewers mentioned problems with the landline concept (elderly and non-English speakers) and with some respondents counting phones, and not distinct landline numbers. One elderly respondent actually started counting the digits in her 10-digit phone number. Some interviewers employed various strategies to avoid using the word "landline" (e.g., like saying "regular phone" or providing the contrast "home phone as opposed to cell phone"). One Spanish interviewer reported that there is no direct translation for the word "landline" in Spanish. There was confusion as to what to do about business lines; minor problems also noted with respect to computer and fax machine references. Some respondents and interviewers thought the question was too wordy. No problems noted with respect to "household" concept.

Assessment of Predictions

Prediction Q1 (A): **Hit**.

Prediction Q1 (B): **Hit**.

VER1. 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?

Prediction VER1 (A): Problems anticipated with clarity (terms, such as "service," "household" and "incoming and outgoing calls"). **Note:** Less than 5% or respondents will be asked this verification item. There may be some annoyance associated with this question, but it should serve its purpose.

Prediction VER1 (B): A small percentage of households that have "telephone service" for internet or fax service only, and this may cause some problems for interviewers.

Evaluation Data/Information

Behavior Coding: Insufficient data (N=2).

Interviewer Debriefing: Of the twelve supplement items rated, VER1 placed eleventh with respect to its average difficulty rating (1.25). Only one log comment was recorded.

Assessment of Predictions

Prediction VER1 (A): **Insufficient data** to evaluate.

Prediction VER1 (B): **Insufficient data** to evaluate.

VER2: I just want to verify that your household has [fill with Q1 response] distinct telephone NUMBERS: Is that correct?

Prediction VER2 (A): Problems anticipated with clarity (terms, such as "distinct," "household" and "telephone numbers").

Prediction VER2 (B): As many as 30-40% of respondents may be asked this verification item and most will have answered Q1 correctly; as an unavoidable consequence, some time-pressed respondents may be annoyed with the follow-up question. Interviewers will not be happy either, but asking this item is necessary if the sponsor wants accurate data.

Evaluation Data/Information

Behavior Coding: There were a relatively small number of cases available for coding purposes (N=10), so these data should be interpreted with caution. Interviewers did a fairly good job reading VER2 as worded (exact code: 80%). In one case, the interviewer simply said something like: "Was that two distinct lines." It would not have been surprising to learn that, after many such cases (i.e., Q1 responses of two or more), interviewers simply probed to determine if the respondent was counting telephones or telephone lines rather than enter a number for Q1, then get the VER2 item, and then have to go back to change Q1 if the earlier response was incorrect. Respondents appear to have little difficult providing an adequate answer (90%), though in 20% of the cases, they did feel the need to clarify their situation (e.g., "One is for the PC, the other for calls.").

Interviewer Debriefing: Of the twelve supplement items rated, VER2 placed eighth with respect to its average difficulty rating (1.56). Three interviewers recorded a total of three comments on their log forms. Two interviewers specifically mentioned that VER2 was useful in catching Q1 response errors (i.e., reporting the number of telephones instead of distinct telephone numbers.). During one of the focus group sessions, one interviewer mentioned that she/he did not think it was necessary to ask the verification items.

Assessment of Predictions

Prediction VER2 (A): **Insufficient data** to evaluate.

Prediction VER2 (B): **Insufficient data** to evaluate.

Q1a. Excluding any numbers used only for faxes and computers, how many of these [fill with Q1 response] landline telephone numbers are used for incoming calls?

Prediction Q1a (A): Problems anticipated with instructions ("excluding" and "only"), clarity (terms, such as "faxes", landlines, "incoming calls"; vagueness; an unspecified reference period) and assumptions (logic, why pay for the phone if you don't take/make calls; different tech arrangements).

Prediction Q1a (B): This item follows VER2. Many respondents are apt to be puzzled by this question, not because it is difficult in a cognitive sense, but rather because the answer would appear obvious: For those households without dedicated PC or fax lines, the answer would be the same number as the number of distinct telephone numbers. Why pay for a phone line if you are not going to take incoming calls?

Prediction Q1a (C): A small percentage of respondents may be uncertain as to how to respond if one or more lines are available for incoming calls but are rarely used (e.g., a PC line versus the household phone number that everybody uses).

Prediction Q1a (D): A small percentage of respondents may interpret the question in an overly exclusive sense. For example, discounting or overlooking the word "only", if a single line is used for incoming calls and for PC internet use, they may respond "no" to this question.

Prediction Q1a (E): A very small percentage of respondents may wonder why the question does not mention outgoing calls and may experience some confusion as a result.

Evaluation Data/Information

Behavior Coding: There were a relatively small number of cases available for coding purposes (N=12), so these data should be interpreted with caution. Interviewer did an excellent job reading Q1a as worded (exact code: 100%). Respondents did a fairly good job providing adequate answers (83%), though there were some signs of confusion. For example, one elder respondent asked the interviewer if she was talking about a personal computer. Another respondent answered that both telephones had the *capability* to take incoming calls, but did not state definitively that they were both used for that purpose.

Interviewer Debriefing: Of the twelve supplement items rated, Q1a placed last/twelfth with respect to its average difficulty rating (1.06). Four interviewers recorded a total of five comments on their log forms, but four of these comments appear to have been intended for Q1b. One interviewer noted some confusion with a respondent who had two landlines, one of which was used exclusively for a fax machine.

Assessment of Predictions

Prediction Q1a (A): **Miss**.

Prediction Q1a (B): **Miss**.

Prediction Q1a (C): **Insufficient data** to evaluate.

Prediction Q1a (D): **Insufficient data** to evaluate.

Prediction Q1a (E): **Insufficient data** to evaluate.

Q1b. Excluding a number used only for a fax or computer, do you [or any other members of your household] take incoming calls on a landline number?

Prediction Q1b (A): Problems anticipated with instructions ("excluding" and "only"), clarity (terms, such as "fax", "landline number", "incoming calls"; vagueness; an unspecified reference period) and assumptions (logic, why pay for the phone if you don't take/make calls; use of filters; different PC/fax/phone arrangements; double barreled). [**Note:** *This question is only asked during personal interviews of respondents who report that they have one landline number.* Many respondents are apt to be puzzled by this question, not because it is difficult in a cognitive sense, but rather because the answer would appear obvious: For those households without a dedicated PC or fax line, the answer would be "one". Why pay for a phone line if you are not going to take incoming calls?]

Prediction Q1b (B): A small percentage of respondents may interpret the question in an overly exclusive sense. For example, discounting or overlooking the word "only", if a single line is used for incoming calls and for PC internet use, they may respond "no" to this question.

Prediction Q1b (C): A very small percentage of respondents may wonder why the question does not mention outgoing calls and may experience some confusion as a result.

Evaluation Data/Information

Behavior Coding: There were a relatively small number of cases available for coding purposes (N=8), so these data should be interpreted with caution. The percentage of exact codes for this item is relatively low (75%), with major wording changes being made 25% of the time. Respondents provided adequate answers (yes/no) 100% of the time. **See note below.**

Interviewer Debriefing: Of the twelve supplement items rated, Q1b placed fourth with respect to its average difficulty rating (2.26). Four interviewers recorded a total of four comments on their log forms, but, as noted above, four comments made with respect to Q1a appear to have been intended for Q1b. Several interviewers mentioned the redundancy of this question when taking the perspective of respondents who had just answered their calls. One interviewer mentioned that she had to explain what a landline was on several occasions; another mentioned an elderly Oriental gentleman having difficult with this item. **See note.**

Note: Item Q1b should not have been asked by interviewers at the telephone centers where behavior coding and interviewer debriefings were conducted. When this question was asked at the telephone centers, it was because interviewers had miscoded a check item ("PORT") regarding the type of interview being conducted, personal versus telephone. Technically speaking, this entry error makes it difficult to evaluate the predictions for this item; as a result, two of the conclusions will be coded as *insufficient data*. The entry error notwithstanding, there simply was not enough negative data or commentary to support the first prediction.

Assessment of Predictions

Prediction Q1b (A): **Miss**.

Prediction Q1b (B): **Insufficient data** to evaluate.

Prediction Q1b (C): **Insufficient data** to evaluate.

Q2. [Excluding students living away at school] Do you [or any other members of your household] have a working cell phone number?

Prediction Q2 (A): Problems anticipated with instructions (e.g., excluding students), clarity (terms such as, "living away," "household," "working"; vagueness), and assumptions (household structure; what does "working" mean; double-barreled).

Prediction Q2 (B): Given that each cell phone has a unique number, it seems odd to be asking about "cell phone numbers" instead of simply asking about "cell phones". This may confuse some respondents.

Prediction Q2 (C): Some respondents may be uncertain as to what the term "working cell phone number" means, especially with respect to the adjective "working"? If someone only makes calls on her cell phone, and rarely if ever has it turned on, is that considered a working cell phone *number*? What if the cell phone is currently "dead" and needs to be recharged: Is that still considered a working cell phone number?

Prediction Q2 (D): Possibly problematic for a household of unrelated individuals.

Evaluation Data/Information

Behavior Coding: Interviewers did a good job reading Q2 as worded (exact code: 90%); however, in about 10% of cases, they did read this question with minor (5%) or major changes (5%) in wording. Respondents, too, performed very well, providing adequate answers 97% of the time; however, the nature of this question was such that respondents felt the need to elaborate upon or clarify their answers about 20% of the time. For example, if there was more than one or more cell phone in the family, the respondent would often elaborate on ownership: "I have one." or "Yes, my wife and I both have one." Responses such as these often have implications for subsequent interviewer behavior: Does the interviewer ignore this additional information about the number of cell phones in the household and just read subsequent questions as worded (standardization view) or does one acknowledge and verify a prior response in lieu of reading a redundant question (flexible-interviewing view)? Though rarely voiced, another example of elaboration was also informative: "Yes, but we don't want to give numbers out."

Interviewer Debriefing: Of the twelve supplement items rated, Q2 placed fifth with respect to its average difficulty rating (1.78). Eight interviewers recorded a total of twelve comments on their log forms. Several respondents were uncertain as to whether their employer-provided cell phones should be counted in addition to their personal cell phones. Others were confused by the term "working", wondering for example if Q2 was specifically about cell phones used at or for work. Others wondered if the cell phone should be considered "working" if it was only being used in emergencies. Some respondents living in childless households (e.g., elderly couples) were disconcerted by the opening phrase "excluding students living away at school". There was also some uncertainty noted with respect to prepaid cell phones (e.g., "trac phones).

Assessment of Predictions

Prediction Q2 (A): **Partial Hit:** Some of the anticipated problems are not reflected in the data (e.g., how respondents interpret "living away" and "household"; assumption regarding household structure).

Prediction Q2 (B): **Miss**.

Prediction Q2 (C): **Hit,** in part for an unanticipated reason; overlap with prediction Q2 (A).

Prediction Q2 (D): **Insufficient data** to evaluate.

Q2a. [Excluding students living away at school] How many different cell phone numbers [(do you have?) or (do the members of your household have?)]

[Skip Instructions: If number equals 1, Go to Q2d. If number equals 2 or more, Go to Q2b.]

Prediction Q2a (A): Problems anticipated with reading (e.g., "working" omitted), instructions (e.g., students; cell phone numbers vs. phones), clarity [wording logic, why ask about numbers; terms; vagueness (e.g., activated vs. under-utilized phones); the reference period "window"]; assumptions and response categories (what to do if household has one cell phone).

Prediction Q2a (B): The adjective "working" does not modify "cell phone number" in this question. This could be a problem for respondents who were uncertain as to the meaning of "working" in Q2—such persons might wonder if non-operable cell phones should be included.

Evaluation Data/Information

Behavior Coding: Interviewers did a fairly good job reading Q2a as worded (85% exact readings); however, in about 13% of cases, they did read this question with minor (8%) or major changes (5%) in wording. Respondents performed very well, providing adequate answers 100% of the time.

Interviewer Debriefing: Of the twelve supplement items rated, Q2 placed ninth with respect to its average difficulty rating (1.50). Three interviewers recorded a total of three comments on their log forms. Two comments were about the irrelevance of the "excluding students" phrase and the third reflected uncertainty with respect to employer-provided cell phones.

Assessment of Predictions

Prediction Q2a (A): **Miss**.

Prediction Q2a (B): **Miss**.

Q2b. How many of the [fill with response to Q2a] cell phone numbers you have do you [or any other members of your household] use regularly?

Prediction Q2b (A): Problems anticipated with reading (e.g., "working" omitted), clarity [terms, such as "regularly"; vagueness (e.g., what does it mean to "use" a cell phone "regularly"); an unspecified reference period], and assumptions.

Prediction Q2b (B): The term "regularly" is undefined for respondents and may pose problems for respondents who use their phones infrequently (e.g., less than once a week most of the time) or on an irregular basis (e.g., some elderly persons).

Evaluation Data/Information

Behavior Coding: Interviewers struggled reading Q2b as worded (exact code: 44%); they read this item with minor wording changes 52% of the time. An awkward four-word sequence ("you have do you") embedded in the middle of the question tripped up most interviewers. In contrast, respondents performed very well, providing adequate answers 96% of the time; however, the nature of this question was such that respondents felt the need to elaborate upon or clarify their answers about 17% of the time. For example, it appeared that in some households where cell phones were used primarily for emergencies, respondents felt the need to explain that use of the cell phone: "Cell phones are not used regularly; for emergencies only."

Interviewer Debriefing: Of the twelve supplement items rated, Q2b placed third with respect to its average difficulty rating (2.38). Ten interviewers recorded a total of fourteen comments on their log forms. Some respondents expressed uncertainty as to the meaning of the term "regularly" (e.g., when cell phone is *used* mostly for emergencies or special occasions); and, as noted above, quite a few interviewers commented on the awkward four-word sequence noted above. Several interviewers also had concerns about the effect of an unspecified reference period on data quality.

Assessment of Predictions

Prediction Q2b (A): **Hit**.

Prediction Q2b (B): **Hit,** overlap with prediction Q2a (A).

Note: Did not anticipate the problems interviewers would have reading this question as worded. Let's classify this as follows: **Miss**.

Q2c. How many of the [fill with the response to Q2a] cell phone numbers are answered by more than one household member?

Prediction Q2c (A): Problems anticipated with clarity [terms, such as "answered"; vagueness (e.g., answer when ringing vs. checking messages vs. filtering; ever?); an unspecified reference period]; assumptions (ignores differences in cell phone usage), and sensitivity/bias (privacy). **Note:** Unlike Q1a or Q1b, which allude to calls received at home, cell phones can be answered at home or away from home; moreover, they tend to be associated with an individual rather than a family. Reference to "other members of your household" in this question (and Q2e) may be equivalent to asking if any of the cell phones noted are *family* (as opposed to *personal*) cell phones.

Prediction Q2c (B): This question follows Q2b if household size is greater than one. The main problem with Q2c may be how literally to interpret the question. Should the respondent answer "yes" if another household member has ever answered a particular cell phone "number" or assume some regularity in this behavior, as specified in Q2b?

Evaluation Data/Information

Behavior Coding: Interviewers were quite successful reading Q2c as worded (exact code: 95%; N=21). And respondents did a fairly good job providing adequate answers (85%; N=20); however, one-quarter of the responses were coded as either inadequate (10%) or "other" (15%). Again, some respondents felt the need to elaborate upon or clarify their use of the cell phones, sometimes not even providing a specific answer to the question: (1) "One is my husband's, one is mine."; (2) "My wife uses one; I have two."; (3) "Well, one's mine and one's my husband's. He answers his and the kids share [mine] some of the time."

Interviewer Debriefing: Of the twelve supplement items rated, Q2c placed second with respect to its average difficulty rating (2.55). Thirteen interviewers recorded a total of nineteen comments on their log forms. Interviewers reported that respondents seemed to have difficulty deciphering the intent of Q2c and determining what sort of answer was required. For example, one interviewer noted that some of her respondents would answer "two" or "both of them", which she recognized to be a fairly rare scenario, so she'd verify by probing: "So you each answer both phones?" And then the respondent would say: "Oh no, he answers his and I answer mine." The interviewers also experienced problems with the response range (i.e., 1-99). In a number of cases, zero was the appropriate answer; and even though it was not included in the specified range, the computer accepted an entry of zero. Some respondents appeared to struggle with the concept of "answered by" wanting to qualify it by considering a frequency component. For example, one respondent answered: "If it's regular, it's by none; but if it's not regular, I have one phone where it's answered by me and my kids."

Assessment of Predictions

Prediction Q2c (A): **Miss**.

Prediction Q2c (B): **Hit**.

Note: Did not anticipate the problems that some respondents would experience in trying to decipher the intent/meaning of this item. Classify as: **Miss**.

Q2d. Do you [or members of your household] regularly answer this cell phone number?

Prediction Q2d (A): Problems anticipated with clarity [wording, such answering phone vs. "number"; terms, such as "regularly" and "answer"; vagueness (e.g., answer when ringing vs. checking messages vs. filtering; ever; an unspecified reference period]; and assumptions (ignores differences in cell phone usage).

Prediction Q2d (B): The term "regularly" is undefined for respondents and may pose problems for respondents who use their phones infrequently (e.g., less than once a week, on average) or on an irregular/emergency basis (e.g., some elderly persons).

Evaluation Data/Information

Behavior Coding: Interviewers did a fairly good job reading Q2d as worded (exact code: 88%; N=17). And respondents did a very good job providing adequate answers (94%; N=16); however, one-quarter of the responses were coded as either inadequate (6%) or "other" (19%). For example, it appeared that in some households where cell phones were used primarily for emergencies, respondents felt the need to explain that use of the cell phone: "Don't give it out; only for emergencies."

Interviewer Debriefing: Of the twelve supplement items rated, Q2d placed tenth with respect to its average difficulty rating (1.40). Eight interviewers recorded a total of nine comments on their log forms. Interviewers reported that respondents seemed to have difficulty deciphering the intent of Q2d and determining what sort of answer was required. For example, one interviewer mentioned the following response: "What does that mean? Do you mean do I let it ring instead of answering it? Another response: "If it's ringing we answer it; but [normally] it's just for emergency purposes." One interviewer alluded to the differences in the way different members of a household use their cell phones: Whereas some adults may have their cell phones off most of the time, all of the teenagers in her extended family have their cell phone 'on' 24-7. The issue of privacy (and the double-barreled structure of Q2d) was subtly captured when one respondent answered this question not with a "yes" response (which would have been both accurate and adequate), but rather by stating that her son had a cell phone and *he* answers it. A "yes" response would have been ambiguous (who actually answers the phone, you or someone else) and could have interpreted by the interviewer that her son's privacy was being violated.

Assessment of Predictions

Prediction Q2d (A): **Partial Hit,** some of the anticipated problems are not reflected in the data (e.g., wording, answering cell phone vs. number).

Prediction Q2d (B): **Hit,** overlaps with Q2d (A).

Q2e. Is this cell phone number answered by more than one household member?

Prediction Q2e (A): Problems anticipated with clarity [terms, such as "answered" and "cell phone *number*"; vagueness (e.g., answer when ringing vs. checking messages vs. filtering; ever; an unspecified reference period]; and assumptions (ignores differences in cell phone usage).

Prediction Q2e (B): This question follows Q2d (an either/or question) for households with more than one member and will seem redundant with that item if the response to Q2d was intended, however implicitly, to communicate information on usage by multiple household members.

Prediction Q2e (C): Another problem with Q2e may be how literally to interpret the question. Should the respondent answer "yes" if another household member has *ever* answered this cell phone "number" or assume some regularity in this behavior, as specified in Q2d?

Evaluation Data/Information

Behavior Coding: There were a relatively small number of cases available for coding purposes (N=9), so these data should be interpreted with caution. Interviewers experienced some difficulty reading Q2e as worded (exact readings, 78%; major wording changes, 22%). Respondents fared a bit better, providing adequate answers 89% of the time.

Interviewer Debriefing: Of the twelve supplement items rated, Q2e placed seventh with respect to its average difficulty rating (1.65). Three interviewers used their log forms to record a total of three comments, two of which related to question meaning. One interviewer wrote: "Some [respondents] seemed confused—had to repeat or explain. A second remarked: "It could be yes [or] no. Had to be clarified." [Note: The group discussion of item Q2e was limited relative to other items for two reasons: (1) It was only identified as problematic during one debriefing session; and (2) the moderator wanted to save time to discuss Q3, an item that clearly was more problematic. Given the similarity in content for items Q2c and Q2e, one might reasonably expect that similar problems would arise for both items.]

Assessment of Predictions

Prediction Q2e (A): **Insufficient data** to evaluate.

Prediction Q2e (B): **Insufficient data** to evaluate.

Prediction Q2e (C): **Insufficient data** to evaluate.

Q3. Of all the phone calls that you [or any other members of your household] receive, about how many are received on a cell phone? Would you say …

<1> All or almost all calls,
<2> More than half,
<3> Less than half, or
<4> Very few or none?

Prediction Q3 (A): Problems anticipated with clarity [wording, due to awkward fills and the reading of response options; terms, such as "all phone calls" and "receive/received"; vagueness (e.g., inclusion/exclusion of personal and business calls; home and away from home]; an unspecified reference period); assumptions (e.g., regularity of behavior by individuals across time); knowledge/memory (e.g., impossible task for large households); response categories (possibly vague/subjective; overlapping; missing "half" of all calls).

Prediction Q3 (B): For large households, especially those with teenagers, this will be a very difficult estimation task. In fact, use of the "or" conjunction makes the task highly ambiguous. For whom is the respondent to report, himself, someone else in the household, or everybody else in the household above the age of 16? If the latter, then the conjunction "and" should have been used.

Prediction Q3 (C): Use of the word "all" in this question is ambiguous: Does it mean all calls received *at home* on a landline (Q1) or a cell phone, or all calls received by family members, anywhere (home or away from home)?

Evaluation Data/Information

Behavior Coding: Interviewers experienced some difficulty reading Q3 as worded (exact code: 73%; minor change code: 23%). 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 (e.g., inadequate answer code: 10%; request for clarification code: 13%; "other" code: 17%). A response of "half" accounted for most of the inadequate answers that were observed; however, there were interesting exceptions: "Different for each of us. Her about none and me more than half." Among the requests for clarification, we heard the following: (1) "Which cell phone?"; (2) "As compared to what?" Among the "other" codes: "In the house, very few or none are received on a cell phone."

Interviewer Debriefing: Of the twelve supplement items rated, Q3 placed first with respect to its average difficulty rating (2.60). Sixteen interviewers recorded a total of twenty-seven comments on their log forms. Interviewers identified a variety of problems with this item: (1) an incomplete response scale (i.e., no "half" option); (2) uncertainties with respect to the response task (e.g., what household members and types of calls to include); and (3) difficulties with respect to the estimation task (e.g., how to generate an estimate of cell phone use in large households).

Assessment of Predictions

Prediction Q3 (A): **Hit,** for the most part; however, some of the anticipated problems are not reflected in the data (e.g., uncertainty caused by the item's unspecified reference period; what might be meant by the terms "receive/received").

Prediction Q3 (B): **Hit,** overlap with prediction Q3 (A).

Prediction Q3 (C): **Hit,** overlap with prediction Q3 (A).

Table 4. Percentage and Frequency of Interviewer and Respondent Behavior Codes for Twelve Supplement Items

Q Label	N	Interviewer Codes[1]				N	Respondent Codes[1]							Comments[2]
		E	mC	MC	PVF		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.

Q Label	N	Interviewer Codes[1]				N	Respondent Codes[1]							Comments[2]
		E	mC	MC	PVF		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)	10% (3)	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,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 5. Difficulty Ratings Assigned to Supplement Items

		Interviewer Number											
Item	**TC**	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	**Mean**	**SD**
Q1	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***
VER1	TTC	*io*	*io*	*io*	2	2	*io*	1	*io*	*io*	1	1.50	*0.577*
	HTC	-	-	-	-	-	-	-	-	-	-	[1.00]*****	-
											Totals:	**1.25**	**-**
VER2	TTC	*io*	*io*	4	4	1	1	3	1	1	1	2.00	*1.414*
	HTC	1	1	*io*	1	2	1	1	1	1	1	1.11	*0.333*
											Totals:	**1.53**	***1.068***
Q1a	TTC	-	-	-	-	-	-	-	-	-	-	[1.00]*****	-
	HTC	1	2	*io*	1	1	1	1	1	1	1	1.11	*0.333*
											Totals:	**1.06**	**-**
Q1b	TTC	*io*	*io*	5	2	4	2	2	2	3	3	2.88	*1.126*
	HTC	2	3	*io*	1	2	*io*	2	1	1	1	1.63	*0.744*
											Totals:	**2.25**	***1.236***
Q2	TTC	2	1	3.5	2	2	*b*	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***
Q2a	TTC	2	1	3	2	2	1	2	1	2	4	2.00	*0.943*
	HTC	-	-	-	-	-	-	-	-	-	-	[1.00]*****	-
											Totals:	**1.50**	**-**
Q2b	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***
Q2c	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***
Q2d	TTC	-	-	-	-	-	-	-	-	-	-	[1.00]*****	-
	HTC	1	3	1	3	2	1	3	2	1	1	1.80	*0.919*
											Totals:	**1.40**	**-**
Q2e	TTC	-	-	-	-	-	-	-	-	-	-	[1.00]*****	-
	HTC	1	4	4	2	3	1	3	3	1	1	2.30	*1.252*
											Totals:	**1.65**	**-**
Q3	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***

[continued on next page]

Question and Scale Used to Rate Problematic Supplement Items:

Based on your experiences this past week, about how frequently did the respondents you interviewed have difficulty providing an adequate answer to this question?

- 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

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.

Notes: (1) TTC interviewer number 3 assigned two precodes to several items which resulted in fractional (average) values for these items. (2) Dashes (-) signify that the item was not identified as problematic by a group of interviewers and therefore was not rated. (3) For those five items not identified as problematic in one of the two focus group sessions, a rating value of 1.00 was assigned for the purpose of computing averages. Those values appear in brackets and are further identified with an asterisk.

Table 6. Inter-Correlation Matrix of Evaluation Techniques

		QAS	HTC Rating	TTC Rating	Avg. Rating	BC 'E'	BC 'Not-E'	BC 'AA'	BC 'not-AA'
QAS	Correlation	1							
	p-value (one tail)								
HTC Rating	Correlation	.387	1						
	p-value (one tail)	.107							
TTC Rating	Correlation	.375	.274	1					
	p-value (one tail)	.115	.194						
Avg. Rating	Correlation	.471	[.711]**	[.871]**	1				
	p-value (one tail)	.061	.005	.000					
BC 'E'	Correlation	-.284	**-.497***	-.357	**-.517***	1			
	p-value (one tail)	.186	.050	.127	.042				
BC 'Not-E'	Correlation	.311	.393	.243	.382	[-.949]**	1		
	p-value (one tail)	.163	.103	.223	.110	.000			
BC 'AA'	Correlation	-.277	**-.526***	-.147	-.376	[-.020]	[.033]	1	
	p-value (one tail)	.192	.039	.324	.114	.475	.460		
BC 'not-AA'	Correlation	.142	**.593***	.091	.370	[-.175]	[.179]	[-.831]**	1
	p-value (one tail)	.330	.021	.389	.118	.293	.289	.000	

Notes: (1) Asterisk (*) indicates correlation is significant at the p=.05 level (one tailed). Parentheses indicate that the correlation is considered non-informative. (2) Of the eight indicators identified in this table, only the QAS represents a pre-survey evaluation method. The other seven indicators are associated with the two post-survey evaluation methods (interviewer debriefing and behavior coding). (3) We correlated three indicators for interviewer ratings associated with problematic supplement questions, the indicator for the Hagerstown group of interviewers (HTC rating), the Tucson group of interviewers (TTC), and the overall average (avg. rating). (4) Regarding the "problematic" behavior codes, the "**BC not-E**" indicator aggregates five interviewer codes (minor change in wording, major change, probe, verify and feedback) and the "**BC not-AA**" indicator aggregates seven respondent codes (qualified answer, inadequate answer, request for clarification, interruption, don't know, refusal and other).

Table 7. Item-Based Correlations between Recoded/Binary Interviewer Codes (ICs) and Respondent Codes (RCs) under Two Analytical Conditions [see Note 1]

	Condition One: IC=0 for E only	Condition Two: IC=0 for E or mC	Comments
Q1	.051 (p=.697) N=60	.188 (p=.150) N=60	Low correlations under both conditions.
Q1a	- - N=12	- - N=12	No variance in ICs.
Q1b	.655 (p=.078) N=8	.655 (p=.078) N=8	ICs are the same under both conditions. Low N.
Q2	-.219 (p=.099) N=58	-.161 (p=.228) N=58	Correlations between ICs and RCs are negative in both conditions.
Q2a	-.087 (p=.595) N=40	-.053 (p=.747) N=40	Correlations between ICs and RCs are negative in both conditions.
Q2b	.462* (p=.026) N=23	.109 (p=.621) N=23	Correlation between ICs and RCs drops dramatically in Condition Two.
Q2c	- - N=20	- - N=20	No variance in ICs.
Q2d	-.149 (p=.582) N=16	- - N=16	Correlation between ICs and RCs is negative in Condition One. No variance in ICs for Condition Two.
Q2e	.655 (p=.078) N=8	.655 (p=.078) N=8	ICs are the same under both conditions. Low N.
Q3	-.262 (p=.162) N=30	-.174 (p=.359) N=30	Correlations between ICs and RCs are negative in both conditions.
VER1	- - N=2	- - N=2	No variance in ICs. Very low N.
VER2	.375 (p=.286) N=10	-.167 (p=.645) N=10	Correlations between ICs and RCs are positive in Condition One and negative in Condition Two.

Note 1: Prior to conducting correlations, interviewer codes [ICs] and respondent codes [RCs] were recoded as "0" or "1". In *Condition One*, an exact question reading [E] was coded as a "0" and all other ICs were recoded as "1". In *Condition Two*, an exact question reading and a minor change [mC] in reading were both coded as "0". For the respondent codes [RCs], an adequate answer [AA] was coded as "0" and all other RCs were recoded as "1".

ATTACHMENTS

Table A-1: LOG FORM Instructions for Focus Group Participants [Cell Phone Use Supplement, February 2004]

INSTRUCTIONS: On the *attached* LOG FORM (tan colored sheets), please keep a *daily written record* of any problems that you or your respondents experience during survey week with regard to the supplement questions listed below. See the LOG FORM for more details.

Please bring the LOG FORM with you when you come to the focus group session. Thank you.

Label	SUPPLEMENT QUESTIONS
Q1	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. How many different landline telephone numbers does your household have?
VER1	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?
VER2	I just want to verify that your household has [fill Q1] distinct telephone NUMBERS: Is that correct?
Q1a	Excluding any numbers used only for faxes and computers, how many of these [fill Q1] landline telephone numbers are used for incoming calls?
Q1b	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?
Q2	[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?

Q2a [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 >1]

Q2b 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?

Q2c How many of the [fill Q2a] cell phone numbers are answered by more than one household member?

Q2d Do you [fill (or members of your household) if NUMHOU > 1] regularly answer this cell phone number?

Q2e Is this cell phone number answered by more than one household member?

Q3 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 …

<1> All or almost all calls,
<2> More than half,
<3> Less than half, or
<4> Very few or none?

INSTRUCTIONS: Please use this log form to identify any supplement questions that are causing problems for you or your respondents—add sheets as needed. Additional information on problem types and sample log entries are provided below for illustrative purposes.

Please include the *question label* associated with the problematic question—for example, Q1, VER2, Q2, Q2b—when describing a particular problem.

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

- ***Coding problems,*** such as when an interviewer is uncertain as to how to code a vague response or has difficulty matching a respondent's answer to available precodes
- ***Comprehension problems***, such as when the respondent has difficulty understanding a particular question or any specific words/terms used in that question
- ***Estimation problems,*** such as when, for example, a respondent has difficulty estimating how many working cell phones the household owns or how frequently they are being used
- ***Proxy problems***, such as when the respondent appears to have difficulty answering a particular question about another household member's actions or behavior

Sample Log Entry: EXAMPLE 1

Q1: The respondent—a sweet elderly woman—did not understand the term "landline". Actually, she thought she heard me say "landmine" initially. We both exploded in laughter when she mentioned that. Had to explain that the phone she was holding, the same one she had been using for decades, is a landline telephone. She had no other telephones, so I entered "1".

Sample Log Entry: EXAMPLE 2

Q3: The respondent, an 18-year old, said the family had maybe six cell phones and three landlines, and that he was in school or working most of the time. He said there was no way he could estimate what percentage of calls was being received on cell phones. I entered "D".

START MAKING ENTRIES HERE:

Table A-2. Debriefing Plan: Overview of Focus Groups Sessions (Cell Phone Use Supplement: February 2004)

1 Introduction of Moderator and Other Guest(s)

2 Purpose of Focus Group Session

3 Description of Focus Group Method and Procedures

4 Brief Introductions of Focus Group Participants [CPS Interviewers]

5 Identify and then Rate Problematic Items [LOGS and Rating Forms]

6 Detailed Discussion of Problematic Supplement Items and Metadata.

7 Discussion of Other Problems with Supplement [Time Permitting]

8 Draw Session to a Close and Collect LOGS.

Note: We will take a 10-minute break at "half-time" to re-energize ourselves.

Table A-3. A Framework Relating Questionnaire Design-and-Evaluation Processes to Sources of Measurement Error

		Interdependent Sources of Measurement Error (at P7)				
		Questionnaire D-and-E Team		Information/Data Collection Context		
	INITIAL DESIGN	*Content Specialists (1)*	*Design Specialists (2)*	*Interviewer (3)*	*Respondent (4)*	*Mode (5)*
P1	*Observation*	C_{11}: ?	C_{12}: ?	▪	▪	
P2	Evaluation	▪	▪	▪	▪	
P3	*Conceptualization*	C_{31}: ?	C_{32}: ?	▪	▪	
P4	Evaluation	▪	▪	▪	▪	
P5	*Operationalization*	C_{51}: ?	C_{52}: ?	▪	▪	▪
P6	Evaluation	C_{61}: A,C	C_{62}: A,C	C_{63}: (A),B	C_{64}: (A)	C_{65}: A
P7	***Administration***	C_{71}: Feb. 2004	C_{72}: Feb. 2004	C_{73}: Feb. 2004	C_{74}: Feb. 2004	C_{75}: Feb. 2004
P8	Evaluation	C_{81}: D,E	C_{82}: D,E	C_{83}: D,E	C_{84}: D,E	C_{85}: D,E

Notes: (1) "No activity" cells, designated with bullet symbols (▪), indicate that no documented activity was conducted or recorded. A question mark (?) suggests uncertainty regarding if and when certain high-probability activities were conducted. (2) With the exception of P7 entries (which note the supplement's administration date), letters represent evaluation methods: **A**=cognitive interviews (June/July 2003), **B**=SORT comments (September 2003); **C**=QAS data (February 2004); **D**=Behavior coding (February 2004); and **E**=interviewer debriefing (February 2004). Parentheses around a letter [e.g., (A)] suggest a nonstandard view of a particular source. See text of paper for information of how the various methods were implemented. (3) Content specialists can be involved in the implementation of a particular method either directly (e.g., in person) or indirectly (e.g., via written documents, like instructional memos or manuals or other metadata).