STUDY SERIES
(Survey Methodology #2009-12)

Person-Based Data Collection in Practice: An Evaluation of Interviewer/Respondent Interactions

Jennifer Hunter
Ashley Landreth

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

This paper was previously published in the 2005 Proceedings of the American Statistical Association.

Report Issued: July 10, 2009

Person-Based Data Collection in Practice: An Evaluation of Interviewer/Respondent Interactions

Jennifer Hunter and Ashley Landreth
U.S. Census Bureau

Abstract[1]

This paper investigates a person-based method of data collection in which a series of demographic questions is asked for each person in the household. Interviewers are instructed to read the questions as worded, using standardized interviewing techniques. However, in field observations it became apparent that the more people that were in the household, the less likely the interviewer was to continue reading the questions as scripted. We are interested in whether or not this practice leads to more respondent behavior problems. To address this, we behavior-coded a sample of interviews using a demographic questionnaire. We analyzed the interviewer/ respondent interactions for the first person in the household and compared that to the interactions for the same questions about additional people in the household. The hypotheses were that when the questions were asked the first time, they would be asked in a more standardized way than when they were asked subsequent times and that deviating from the standardized wording would lead to more respondent behavior problems. We report differences in both interviewer and respondent behavior that stemmed from using a more conversational method of interviewing for later persons in the household.

Keywords: Behavior coding, Person-based interview, Pretesting, Interviewer-respondent interactions, Standardized interviewing

1. Introduction

Often survey researchers make the assumption that interviewer-administered surveys will be administered the way they are scripted, particularly when interviewers are trained in standardized interviewing techniques. However, different data collection strategies may impact interviewers' ability to stick to the interview script. When facets of the interview cause interviewers to go off-script, the interview may resemble conversational interviewing. Although conversational interviewing techniques have been studied and are recommended in some situations (see Conrad and Schober, 2000, and Schober and Conrad, 1997), interviewers normally receive training in such techniques prior to implementing them. When an interviewer spontaneously changes the script of a standardized questionnaire, the interviewer may turn a standardized interview into a conversational one without having the proper training. We expect that this would lead to decreased data quality.

This paper investigates the effect of person-based interviewing on interviewer behavior within the context of a demographic census. In person-based interviewing, the respondent is asked a series of questions about the first person in the household before the interview cycles back through the same series of questions about the next member of the household (and so on until all data about each household member have been collected). Interviewers are instructed to read the questions as worded, using standardized interviewing techniques. However, in field observations it became apparent that when household size increased, interviewers were less likely to continue reading the questions as scripted[2]. We are interested in how often this happened in the field and whether or not this practice led to more response problems.

Before we address the study at hand, we recognize that the evaluation presented in this paper hinges on the assumption that standardized interviewing is a desirable goal. There has been considerable discussion of standardized versus conversational interviewing in the survey research literature (see Beatty, 1995). We will take a look at this debate, while examining factors relevant to our data collection needs. Next, we will look for evidence in the literature of the effects of person- versus topic-based administration on interviewing and data quality. Finally, we will present our study that examines the effect of person-based interviewing on the interaction between interviewers and respondents and, presumably, data quality.

[1]This report is released to inform interested parties of research and to encourage discussion. The views expressed on methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau. This paper was presented at the Annual Conference of the American Association for Public Opinion Research, May 15, 2005.

[2] The same questions are asked about each person in a household; therefore, the length of interview and redundancy of content increases with household size.

1.1 Standardized Interviewing

Standardized interviewing reflects the goal that a question should be administered in exactly the same way to each respondent in order to achieve data that are comparable across respondents (see Fowler, 1992; Fowler and Mangione, 1990). Recently, there has been a debate surrounding the benefits of standardized interviewing and whether this goal is realistic, or even desirable (see Beatty, 1995; Suchman and Jordon, 1990). Conrad and Schober (1999) have been major proponents of a type of interviewing that contrasts with traditional standardized interviewing. They have conducted several studies, both in the lab and in the field that show the benefits of a more conversational type of interviewing where interviewers are given the liberty to explain concepts to respondents and probe until they come to a mutually agreed-upon answer with the respondent (e.g., Conrad and Schober, 1996; Schober and Conrad, 1997). They have found that for complicated situations, conversational interviewers elicited better data than standardized interviewers did. However, they found no difference in data quality between the methods for easy situations. The authors point out that the drawback for conversational interviewing is the increased cost associated with increased interviewing time.

While there may be a place for conversational interviewing, there are several reasons why we believe that the census is not one of them. Beatty (1995) mentions a key question that comes up in the standardized/conversational debate: "Should interviewers be 'knowledgeable experts'" . . . or . . . "non-experts who obtained information on a mass-scale using standardized techniques?" (pg. 148). The answer to this question certainly depends on the survey about which one is interested. In this paper, we are addressing, in particular, the collection of decennial census data.[3] By nature, our interviewers are predominantly non-experts (i.e., employees who are hired only for a period of a few months surrounding the census), who must collect information on a mass-scale (i.e., about 20% of all census data are gathered through these, mostly novice, interviewers in the non-response follow-up operation; Treat, 2004). Additionally, Conrad and Schober have commented that standardized interviewing works just as well as conversational interviewing when "concepts in the question clearly correspond to the respondent's life circumstances" (1996; pg. 883). Although the demographic questions asked on the census short form do pose problems for some people, the general concepts of age and gender are not among the most difficult topics for which survey designers collect data. For these reasons, we believe that it is justifiable to evaluate the quality of our census interviews using standardized interviewing as the goal.

1.2 Person- vs. Topic-based Administration

There are two basic approaches for administering a series of questions where data are collected from one respondent about multiple people. The first is a person-based approach, which consists of a series of questions that are asked in their entirety about the first person, then the same series is administered again about the next person, and so on (e.g., sex, age, date of birth, and race data are gathered about Person 1, and then data on the entire series are gathered about Person 2). The alternative method of administration is topic-based, in which data regarding a single topic are gathered for everyone in the household before moving on to the next topic in the survey (e.g., race is gathered for everyone in the household, then age is gathered for everyone in the household).

Research suggests that topic-based interviewing may be more "respondent friendly" than person-based interviewing. Moore and Moyer (2002) examined differences in the data gathered by person- and topic-based versions of the American Community Survey (which asks a rather long series of demographic questions about each household member). The person-based approach had lower response rates, higher refusal rates, and longer interview length (see also Colosi, 2001). Interviewers reported that the topic-based approach facilitated rapport building with the respondents and was more helpful with reluctant respondents than the person-based one. Item non-response was higher for the question on race in the person-based instrument than in the topic-based one; that question was the only census (short form) item that had a non-response rate of 2 percent or higher.[4]

In assessing data quality, research suggests that there is no difference in the data obtained through a person-based versus topic-based approach for simple demographic data. Moore and Moyer (2002) investigated the theory that the topic-based approach would lead to more household homogeneity, that is, members of the household would look like they share more characteristics than those given the person-based approach. They found no difference in the proportion

[3] Interviewer-administered interviews are conducted with people who do not respond to the initial mailout phase of the census.

[4] Loomis (1999) did find increased non-response for income items in the American Community Survey using a topic-based approach. However, income information is not collected on the census short form.

of households in which all members shared the same race or Hispanic origin.[5]

All in all, research indicates that there may be some disadvantages to using a person-based instrument. This paper will attempt to further determine what problems exist with the person-based approach by examining interviewer and respondent behaviors during this type of question administration.

1.3 Objective

With standardized interviewing as the preferred tool for collecting census data, the behavior coding method is well suited for evaluating the survey questions. As part of an evaluation of a census non-response follow-up (NRFU) operation, we used behavior coding to analyze the effect of person-based interviewing on interviewer and respondent behavior. In the NRFU, the U.S. Census Bureau tries to collect basic demographic data using an interviewer-administered, almost fully scripted survey instrument, in hopes of producing a standardized interview. The design of this data collection entails person-based interviewing to collect person-level data. The objective of this paper is to determine whether the person-based data collection methodology encouraged standardized interviewing procedures by looking at how interviewers read the questions. The hypotheses, based on anecdotal field observations, were that when the questions were asked the first time, they would be asked in a more standardized way than when they were asked subsequent times and that deviating from the standardized wording would lead to more response problems.

2. Method

2.1 Study Design

The Census Bureau conducts many small-scale field tests throughout the decade to prepare for the next decennial census. As a part of the 2004 Census Test, the Census Bureau conducted a test of the NRFU operation, which is an attempt to gather census data from people who did not respond to the mail-out census form. In 2004, the NRFU survey was a mobile, computer-assisted personal interview (MCAPI), which contained both English and Spanish language questions. Although interviews were conducted in both languages, for this paper we will collapse across language and examine the effect of asking the same questions repeatedly in a person-based manner (referred to as repeated question administration).[6]

The NRFU interview is relatively short (i.e., 7 to 30 minutes per household depending on the number of household members). It begins with household-level questions, and about mid-way through the interview, it switches to person-level questions. This paper focuses on the person-based administration of person-level questions.

Face-to-face interviews were audiotaped throughout the field period of NRFU, from May to July of 2004 in the Queens, NY test site. Of the 256 audiotapes collected, a total of 220 were deemed usable. Though the sample was not designed to be statistically representative, we did achieve a sufficient sample to analyze the questions of interest.

The behavior-coding method is used in survey research to analyze the interactions between interviewers and respondents during the administration of survey questions (Cannell, Fowler, and Marquis, 1968). The method involves the systematic application of codes to behaviors (in this case, verbal behavior) that interviewers and respondents display during the question/answer process, and is often used to identify problematic questions (Oksenberg, Cannell, and Kalton, 1991; Sykes and Morton-Williams, 1987). In an ideal interaction between an interviewer and a respondent, the interviewer asks the question exactly as worded and the respondent immediately provides a response that is easily classified into one of the existing response categories. When the interaction repeatedly deviates from this ideal, we begin to suspect there may be problems with the question and/or response options. The application and analysis of behavioral codes allow researchers to pinpoint where such issues are occurring in the survey instrument (Fowler and Cannell, 1996).

This paper presents results from codes that were designed to capture three main aspects of behavior that occur for each question: 1) question-asking behavior for interviewers; 2) immediate response behavior for respondents (i.e., first-level exchange); and 3) interruptions by respondents (i.e., "break-ins"). The framework of behavioral codes used for this study was adapted from Oksenberg, Cannell, and Kalton's (1991) research and can be found in Hunter and Landreth (2005). In addition to recording the codes themselves,

[5] Among the variables that did show differences in within household homogeneity, there was no consistent direction of the effect. For one variable, person-based interviewing had more homogeneity, for two variables there was more homogeneity for the topic-based instrument, and for three variables (including race and Hispanic origin) there was no effect.

[6] The logistic regression revealed that there were no interaction effects between language and repeated administration. The effects of repeated administration were the same whether or not language was controlled for in the model (see Hunter and Landreth, 2005).

when non-ideal interactions occurred, coders were instructed to transcribe or summarize the verbal interaction for qualitative analysis.

Five telephone interviewers from a Census Bureau telephone center were trained in project-specific behavior-coding techniques and served as the behavior coders. Audio-taped interviews were distributed among coders and each coder coded approximately 50 tapes. To assign codes, they listened to the audiotapes and followed along with a written guide that presented the questions in both languages. Coders made these assessments based upon the audiotapes only; they did not have access to data generated by each interview.

To assess the quality of the behavior coding data, all interviewers independently coded a test set of eight interviews and an inter-coder reliability statistic was generated (a Kappa score). The Kappa statistic provides a conservative measure of agreement among coders in their application of the behavior codes, because it adjusts for the possibility of agreement by chance (Fleiss, 1981). According to Fleiss, a Kappa score between 0.7 and 0.48 represents a good to fair level of agreement. Our average kappa statistic was 0.49, indicating a fair level of agreement.[7] For more information on the inter-coder reliability analysis, see Hunter and Landreth (2005).

For this study, we considered "good" interviewer behavior to include exact wording or slight changes to question wording that did not affect the question's meaning, and correct verification (i.e., when the interviewer correctly verifies information that the respondent gave earlier in the interview). The only behavior that was considered good respondent behavior was an adequate (or codeable) answer. We examined the data by repeated administration (i.e., the first administration of the question in a household versus all subsequent administrations of the same question). We looked at the effect of repeated administration on good interviewer and respondent behavior and respondent interruptions using logistic regression analysis.

2.2 Limitations

Aspects of the research design present limitations to this study and necessitate some caution in interpreting and understanding the results. Audio recording restricts observable behavior to verbal communication, which excludes nonverbal communication that occurs naturally as part of the face-to-face interviewing process. For instance, a respondent might nod his or her head to a yes/no question, but this silent behavior goes undetected on an audiotape. This respondent's behavior would be recorded as "inaudible" (which is coded as opposed to adequate), and therefore the number of adequate answers provided by respondents for a given question may be artificially decreased in the analysis.

Additionally, the results of the statistical tests performed for this study are intended to be used for heuristic purposes only. The tests were performed as if the data were collected with a simple random sample, with replacement. However, these data were not collected randomly; and therefore, the results are not generalizable.

3. Results

Using the behavior coding data, we are able to calculate the percent of good interviewer behavior, respondent behavior and respondent interruption for the first administration of each question and for subsequent administrations. In this section, we first present general findings for interviewer and respondent behavior, and then we explore the effects of repeated administration at the individual question level for the questions that had significant differences in interviewer behavior between first and subsequent administrations.

3.1 Overall Results

3.1.1 General interviewer behavior

We were interested in addressing whether or not interviewers achieved an acceptable level of standardized interviewing behavior using a person-based data collection strategy. Unfortunately, the survey instrument produced an interview that was less standardized, overall, than we had hoped. The commonly accepted error threshold suggests non-ideal interviewer behavior (e.g., major changes to question wording, omitting a question) should occur no more than 15 percent of the time (Oksenberg, et. al, 1991; Fowler, 1992). Table 1 presents the proportions of good interviewer, respondent behaviors, and respondent interruptions by first administration contrasted to later administrations. Across all seven person-level questions, on average, good question-asking behavior across the person-level questions was only 36 percent.

The trend for every person-level question we analyzed was that it was asked as intended (i.e., exactly as worded, with slight changes or correctly verified) more often the first time it was administered than for repeated administrations (i.e., it was asked appropriately more often for Person 1 than for Persons

[7] The Kappa scores were as follows: interviewer behavior (0.57), first-level response behavior (0.41), and final response outcome behavior (0.51).

2 and later in the household; see Table 1).[8] For the first administration, questions were asked as intended on average 47 percent of the time; for subsequent administrations, correct question-asking behavior decreased to 31 percent (see Table 1). In the logistic regression analysis, we found significant effects of repeated question administrations on "good" interviewer behavior for the questions on age, Hispanic origin, and race, in addition to a question attempting to determine if each person could have been counted at another place (this question is called Coverage; see Table 1). These effects will be discussed in further detail during the question-level analysis below.

Based on results of another behavior coding study, Stanley (1996) posited that the prospect of violating conversational norms leads to incorrect interviewer behavior – e.g., asking questions when you already have the answer violates a conversational norm. We suspect that a large portion of the problem with interviewers incorrectly administering the questions for Persons 2 and later is because interviewers feel like they already have the information needed to answer those questions, either because the respondent has already explicitly told them or because the interviewer assumes that the same response applies to all household members. For example, when the interviewer asks the question on race about the first person in the household, the respondent may tell them that all people in the household are of the same race (in anticipation of later questions) or the interviewer may assume that all household members are of the same race as Person 1. We have tried to take into account the former possibility by allowing a correct verification (i.e., when the interviewer correctly verifies information that the respondent gave earlier in the interview) to be included as "good" interviewer behavior. However, there remains a discrepancy between good interviewer behavior the first time a question is administered and subsequent administrations that indicates a problem beyond that of simply not repeating a question to which the interviewer already has an explicit answer.

3.1.2 General respondent behavior

In the logistic regression analysis, we found significant effects of repeated question administrations on good respondent behavior for the Hispanic origin and associated follow-up questions (see Table 1). When respondent behavior was classified as "good" at a higher rate on the first administration than subsequent administrations, it was due to an increased number of inaudible responses for questions about subsequent persons in the household rather than increased inadequate responses. As the interview progresses, respondents may become more likely to nod or shake their heads to answer yes/no questions rather than verbalizing an answer. This does not necessarily indicate a problem with the question/answer process, but rather seems to be an artifact of using audiotapes to code behavior. Thus, we will not discuss these effects any further.

There were no significant effects of repeated administration on respondent interruptions. Rates of respondent interruptions are presented in Table 1 to illustrate the relative magnitude of interruptions.

3.2 Question-level Analysis

In this section, we further explore the questions with percentages of good interviewer behavior that differed significantly between first and subsequent administrations. Behavior-coding allowed us to see problems that we can reasonably presume to be related to two aspects of interviewer behavior: 1) interviewers compensate for what they believe to have established as common ground; and 2) interviewers compensate for questions that they seem to feel are overly-burdensome.

3.2.1 Age question

What was [your/ NAME's] age on April 1, 2004?

The age question was asked more often as intended for Person 1 than for subsequent persons in the household (62% versus 45%; see Table 1). This seemed to be due to a difference in the percent of major change to question wording for Person 1 compared to subsequent persons (33% versus 50%, respectively, Hunter and Landreth, 2005). From the behavior coding notes, it seems that for Person 2 and later, the interviewer was more likely to omit the date. It is possible the interviewers thought the reference date had been established after explicitly stating it for Person 1, so they omitted it for subsequent household members to avoid redundancy. Interviewers may have compensated for what they believed was established common ground (e.g., all questions are asked in relation to April 1, 2004).

3.2.2 Hispanic origin question

[Are you / Is NAME] of Spanish, Hispanic or Latino origin?

[8] This is a significant effect when using a sign test that examines the trend that first administration yields a higher rate of correct interviewer behavior than later administrations (7/7 pairs are like signed; p=.016; see Snedecor & Cochran, 1967).

Overall, the Hispanic origin question was asked correctly only 37 percent of the time (Hunter and Landreth, 2005). This did differ by repeated administration (see Table 1). It was asked correctly 62 percent of the time for Person 1 and only 29 percent of the time for subsequent persons in the household. This question was skipped only 3 percent of the time for Person 1, but it was skipped 27 percent of the time for subsequent people. Correct verifications were also slightly higher for subsequent persons in the household than for Person 1 (0.5% for Person 1 and 5.8% for subsequent persons). This indicates interviewers may have used the response for Person 1 and applied it to all household members, either because the respondent offered that all household members were of the same ethnicity (as the case may be for correct verifications), or because interviewers assumed so (as may have been the case for omissions). Once again, we see an indication that interviewers are not asking questions when they believe that the information has been established previously in the interview.

3.2.3 Race question

Using this list, please choose one or more races that [you / NAME] consider(s) [yourself / himself / herself] to be.

Overall the race question was asked correctly only 20 percent of the time (Hunter and Landreth, 2005). There was an effect of repeated administration on the correct administration of this question as well (see Table 1). For Person 1, it was asked correctly 41 percent of the time; for subsequent household members, it was asked correctly only 15 percent of the time. This question was skipped quite often – 40 percent of the time overall; it was skipped 16 percent of the time for Person 1 and 49 percent of the time for Person 2 and later. Presumably this indicates the tendency for the interviewer to input the same race for all household members (either because the interviewer assumes that this is the case, or because the respondent told them so earlier in the interview, and the interviewer failed to verify it for each person). Overall, 18 percent of the time the interviewer verified race (i.e., 10 percent verification for Person 1's race and 20 percent verification of Person 2 and later's race). This further supports the hypothesis that the interviewer uses Person 1's answer to the race question to infer information about the race of other household members.

3.2.4 Coverage question

The Coverage question is used to help determine if anyone stayed at another place around the time of the census. Interviewers were instructed to read the response set in its entirety.

[Do you / Does NAME / Did NAME] sometimes live or stay somewhere else?

> **To attend college?**
> **To stay at a seasonal or second residence?**
> **To be closer to work?**
> **For a child custody arrangement?**
> **For any other reasons?**

Correct administrations of this question were extremely infrequent (16%); it was more often read with major changes (66%, see Hunter and Landreth, 2005). The first time this question was asked in each household, good interviewer behavior occurred 27 percent of the time, but dropped to 12 percent for all subsequent administrations of this question (see Table 1). The main shift in behavior attributable to this finding seems to be interviewers' increased tendency to skip this question for Person 2 and later. Interviewers skipped the question entirely for 18 percent of all persons, which was only 6 percent of the time for Person 1, but 22 percent of the time for subsequent household members. This seems to indicate one of three possible problems: 1) interviewers received information earlier in the interview that the same answer applied to all household members (and failed to verify it for each person); 2) interviewers assumed the same answer would apply; or 3) interviewers thought the question was overly burdensome.

Respondents' reactions to this question may have caused interviewers to take shortcuts with question wording, if they were not already doing so. Table 1 shows respondent interruptions were greatest for this question compared to all other questions (13% overall). When interruptions occurred the first time this question was posed, it likely encouraged interviewers to change or skip this question for subsequent administrations. In addition to the assumption that the same information could apply to all household members, the complexity of this question likely caused it to be skipped. Interviewers may have learned during the first administration that this question was burdensome, and decided to skip it for subsequent administrations.

4. Conclusions

Standardized interviewing was not achieved in this study using a person-based instrument where the questions were scripted such that an entire battery of questions was asked about a single person in the household before the interview proceeded to collect data about the next person. Interviewers were better

able to stick to the script for the first administration of the question, but they altered or skipped the question more often for subsequent people in the household. This indicates that they either already gathered the information for the subsequent household members (and did not verify it), or they thought it was overly burdensome to repeat the exact same questions again, so they modified them. It is possible that respondents, in anticipation of the survey's intent, began offering relevant information for the entire household at particular questions, perhaps causing interviewers' behavior to deviate from standardized interviewing procedures later in the interview.

If an interviewer thinks common ground has been established, he or she may alter the interview in order to maintain conversational norms and take into account information that has already been provided. This may be more difficult in a person-based interview, because for each person in the household, the interviewer must complete the entire series. Previous survey design research has indicated that a much more natural flow can be achieved by asking questions in a topic-based approach where the interviewer asks, for example, for the age of Person 1, then says "how about Person 2?", "how about Person 3?", and so on through the list of household members (see Colosi, 2001; Fuchs, 1999; Moore and Moyer, 2002). This allows the interviewer to carry on the "conversation" using standardized interview probes. If interviewers spontaneously used this approach with a person-based instrument in a CAPI environment, like the one examined here, it would present a problem because data cannot be recorded in a topic-based manner. Fuchs (1999) notes this problem saying that it causes the interviewer to have to memorize the answers to questions that will be asked about other household members later on in the interview.

The results of the current study suggest that for relatively straightforward, interviewer-administered demographic data collections, topic-based interviewing may better accommodate the respondents' natural tendency to provide certain types of information for the entire household and facilitate interviewers' data capture needs. In addition, it may also prevent interviewers from delivering a nonstandardized interview due to perceived question redundancy.

5. Future Research

The next logical step, though we have no immediate plans to conduct such a study, would be to conduct a split panel field test with half of the administrations being person-based and half being topic-based. Such a study could examine data quality (measured by comparing the distributions of data and the item non-response rates resulting from the different administrations, similar to that done by Moore and Moyer, 2002) as it related to how the interview was administered (behavior coding rates of standardized interviewing for each kind of administration). This would give us more information on person- versus topic-based administration and how standardized interviewing affects data quality in each approach. This would also give a definitive answer as to whether topic-based interviewing would be applied in a more standardized fashion than person-based interviewing.

Acknowledgements

The authors acknowledge Courtney Reiser, Aref Dajani, Pamela Ferrari, Patricia Goerman, LaToya Barnett, Maria Cantwell, Juanita Rasmann, Theresa DeMaio, Elizabeth Murphy, Jeffery Moore, and Jennifer Rothgeb for their technical assistance and comments on an earlier draft of the paper.

References

Beatty, P. (1995). Understanding the standardized/ non-standardized interviewing controversy. *Journal of Official Statistics, 11* (2); 147-160.

Colosi, R. (2001). An Analysis of Time Stamp Data: Person-based Interviewing vs. Topic-based Interviewing. Unpublished report.

Conrad, F. G., and Schober, M. F. (1996). How interviewers' conversational flexibility affects the accuracy of survey data. Paper Presented at the Annual Meetings of the Section on Survey Research Methods of the American Statistical Association.

Conrad, F. G., and Schober, M. F. (1999). Conversational interviewing and data quality. Paper Presented at the Federal Committee on Statistical Methodology Conference.

Conrad, F.G., and Schober, M.F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*, 1-28.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions.* John Wiley & Sons, New York.

Fowler, F. (1992). How unclear terms affect survey data. *Public Opinion Quarterly, 56*: 218-231.

Fowler, F., and Cannell, C. (1996) Using Behavioral 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 Fransisco, CA: Jossey-Bass.

Fuchs, M. (1999). Screen design and question order in a CAI instrument effects on interviewers and respondents. Paper Presented at the Annual

Meetings of the Section on Survey Research Methods of the American Statistical Association.
Hunter, J.E. and Landreth, A. D. (2005). "Behavior coding analysis report: Evaluating bilingual versions of the non-response follow-up (NRFU) for the 2004 Census Test." Available on the SRD Research Report Series http://www.census.gov/srd/www/byname.html
Loomis, L. (1999). An Analysis of Nonresponse to Income Questions in the ACS/CATI Person-Based/Topic-Based Field Experiment. Unpublished U.S. Census Bureau report, January 27, 1999.
Moore, J. C., and Moyer, L. (2002). ACS/CATI Person-Based/Topic-Based Field Experiment. Statistical Research Division Research Report Series, #2002-04.
Oksenberg, L., Cannell, C., and Kalton, G. (1991) New Strategies for Pretesting Survey Question. *Journal of Official Statistics 7* (3): 349-394.
Schober, M. F., and Conrad, F. G. (1997). Does conversational interviewing improve survey data quality beyond the laboratory? Paper Presented at the Annual Meetings of the Section on Survey Research Methods of the American Statistical Association.
Stanley, J. S. (1996). Standardizing interviewer behavior based on the results of behavior coding. Paper Presented at the Annual Meetings of the Section on Survey Research Methods of the American Statistical Association.
Suchman, L., and Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association* 85:232-241.
Snedecor, G.W., and Cochran, W.G. (1967). *Statistical Methods.* Ames Iowa: Iowa State UP.
Sykes, W., and Morton-Williams, J. (1987). Evaluating Survey Questions. *Journal of Official Statistics 3* (2): 191-207.
Treat, J. B. (2004). *Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No. 11, TR-11, Response Rates and Behavior Analysis,* U.S. Census Bureau; March 2004.

Table 1. Percent Good[1] Interviewer and Respondent Behavior and Respondent Interruptions by Repeated Question Administrations

	Good interviewer behavior		Good respondent behavior		Respondent interruptions	
Question	Person 1	Person 2+	Person 1	Person 2+	Person 1	Person 2+
Sex	56.4%	50.8%	59.1%	58.4%	2.5%	3.3%
Age	61.8*	45.4	70.0	71.0	0.5	2.5
Date of Birth	48.3	43.6	77.5	70.6	1.4	1.8
Hispanic	61.7*	29.3	82.3*	69.0	2.9	3.7
Hispanic Follow-Up	34.8	34.5	87.6*	68.0	15.2	9.1
Race	40.8*	14.6	44.1	41.5	1.6	3.8
Coverage	26.8*	11.9	84.8	85.7	16.2	11.6
Average	47.2	31.5	63	67.1	5.8	5.1

* Significant difference[2] at $p < .002$ from the logistic regression analysis.[3]

[1] Exact wording/slight change and correct verification were considered "good" interviewer behavior. The only behavior that was considered good respondent behavior was an adequate answer. Findings were very similar when only exact wording/slight change was used as good interviewer behavior.

[2] The results of the statistical tests performed for this study are intended to be used for heuristic purposes only. The tests were performed as if the data were collected in a simple random sample, without replacement, which was not true in this case.

[3] We conducted a total of 21 tests for this study (7 questions and 3 dependent measures). To ensure a study-wide significance level of .05, we recommend using a Bonferroni adjustment, which lead to a significance level of $p < .002$ (see http://home.clara.net/sisa/bonfer.htm to replicate this adjustment).