Journal of Official Statistics, Vol. 26, No. 3, 2010, pp. 535–557

Bilingual Questionnaire Evaluation and Development through Mixed Pretesting Methods: The Case of the U.S. Census Nonresponse Followup Instrument

Jennifer Childs[1] *and Patricia Goerman*[1]

The objective of this research was to develop and improve a Nonresponse Followup (NRFU) instrument for the U.S. Census. This research is unique because multiple pretesting methods were used in the development of an instrument in two different languages: English and Spanish. This article discusses results of three rounds of English cognitive testing, two rounds of Spanish cognitive testing, two rounds of behavior coding of the instrument in both languages, and an observational study in the field in both languages. The application of mixed pretesting methods to the development of one survey instrument is an all-too-uncommon situation. This article presents lessons learned about the types of findings made possible by the different pretesting methods, and offers the unique opportunity to examine issues of equivalency between a source and a translated version of a survey instrument through multiple measures.

Key words: Bilingual questionnaire development; pretesting methods; cognitive interviewing; behavior coding; observational study.

1. Background

Pretesting of bi- or multilingual survey instruments has recently become an established practice at the U.S. Census Bureau and many other large survey organizations (e.g., Carrasco 2003; Goerman 2006; Harkness 2004; Pan 2004; Willis 2004). In 2004, the U.S. Census Bureau released translation guidelines that recommend pretesting all survey translations for "semantic, conceptual, and normative equivalence" (U.S. Census Bureau 2004). Additionally, the Census Bureau Standard for Pretesting Questionnaires and Related Materials for Surveys and Censuses (U.S. Census Bureau 2003) requires that survey questions be pretested and shown to "work" prior to being fielded. The U.S. Census Bureau standards and guidelines recommend pretesting questions in the languages in which they will be administered. The objective of this article is to use the decennial census Nonresponse Followup (NRFU) instrument as a case study to examine the benefits of using mixed methods of pretesting in the development of a bilingual (English/Spanish)

[1] U.S. Census Bureau, Center for Survey Methods Research, Statistical Research Division, Washington, DC 20233, U.S.A. Emails: jennifer.hunter.childs@census.gov and patricia.l.goerman@census.gov
Acknowledgments: This report is released to inform interested parties of research and to encourage discussion. Any views expressed on the methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau. This article was presented at the Survey Methods in Multinational, Multiregional and Multicultural Contexts (3MC) in Berlin, Germany on June 27, 2008. We would like to thank Theresa DeMaio, Yuling Pan, Stephanie Sheffield, and Nathan Jurgenson for reviewing earlier drafts of this article and providing insightful comments.

survey instrument. This case study shows the different types of results made possible through the application of different pretesting methods to the same bilingual survey instrument. The NRFU instrument was tested through usability testing, cognitive testing, behavior coding, an observational study and large-scale field tests. While the timing and sequencing of the different studies presented here was not ideal, examining the instrument's overall course of development allows us to examine the types of findings made possible by the different pretesting methods. In addition we are able to recommend a more ideal sequence of testing for the future.

As a part of the decennial census operations, the U.S. Census Bureau mails out census forms to most housing units in the country. The U.S. Census Bureau attempts to send an interviewer to every housing unit that does not return a census form by mail. The interviewer asks the household to participate in the census via an in-person interview. This personal visit is a part of the NRFU operation. In preparation for the 2010 Census, the U.S. Census Bureau developed self-administered paper census forms to be mailed to respondents and the Computer Assisted Personal Interview (CAPI) NRFU instrument in both English and Spanish. The development and testing of the bilingual CAPI instrument is the focus of this article. The U.S. Census Bureau had originally planned to collect NRFU data using a CAPI instrument in 2010. Due to a change in plans, the 2010 NRFU data was collected via an interviewer administered paper-and-pencil instrument. Nevertheless, this article discusses lessons learned from the CAPI instrument development process, which will inform future U.S. Census Bureau initiatives.

In the development cycle, the self-administered census questionnaire that was mailed to respondents was created first. The adaptation of this self-administered questionnaire to the CAPI mode necessitated changes in the question wording and administration in order to optimize interviewer and respondent interactions (for historical context of moving from a Paper and Pencil to a CAPI instrument, see Nicholls and de Leeuw 1996; for recent discussion of this topic particularly related to this case study, see Childs and Landreth 2006; for more general discussions of adapting questions for modes see Dillman and Christian 2005; and Martin et al. 2007). The survey questions as adapted for a CAPI instrument are the focus of the pretesting efforts discussed in this article.

The United States census collects very basic data on each housing unit (e.g., whether the unit is occupied or not, whether the unit is owned or rented) as well as some basic demographic data about each person who lives in the household (e.g., names, ages, races). The NRFU instrument uses flashcards (also known as showcards) to assist respondents in answering particularly long or complex questions. Flashcards are used to present instructions on who to count in the household, the relationship between the householder and other residents, and the various origin and race response categories included in the survey instrument.

According to many researchers, when a survey instrument is to be administered in multiple languages, parallel development of the different language versions of the questionnaire, rather than translation of a survey instrument, is ideal (Harkness et al. 2003; Potaka and Cochrane 2004). When parallel development is not possible, it is often recommended that instruments be translated using a team or committee approach (U.S. Census Bureau 2004).

In the case of the NRFU instrument, the instrument was developed first in English and was then translated into Spanish. The translation of the instrument did not involve a formal

committee approach or peer review. Additionally, the programmers who programmed the Spanish instrument onto the handheld computer were monolingual English speakers and therefore it was easy for them to introduce or overlook spelling or grammatical errors in the Spanish version. Unfortunately, there was not a review of the Spanish version of the questionnaire in between the programming and the fielding of the instrument for the field test. While we do not consider this to be the ideal way to develop a multilingual survey instrument, both cost and staffing resources influenced the development process. Pretesting of the wording began only after the questionnaire had been developed in English, translated into Spanish and programmed into the CAPI instrument in both languages. This undoubtedly impacted our findings.

1.1. Cognitive Testing

Cognitive interviewing is a well-known and commonly-practiced form of pretest where social scientists usually conduct a semi-scripted interview with individual respondents with the goal of understanding how respondents comprehend and answer the survey questions (see DeMaio and Rothgeb 1996; U.S. Census Bureau 2003; Willis 2005, for a more detailed explanation of the cognitive interview). Interviewers probe respondents – either concurrently with the administration of the survey questions, or retrospectively, after the survey itself has been completed – to assess how well they understood the questions and concepts being measured as well as the accuracy of their responses, given their personal situations. Some researchers also include "think aloud" protocols in which respondents are asked to think out loud while filling out a questionnaire or while deciding how they will answer questions in an interviewer administered survey (Beatty 2004; Willis 2005). Results from cognitive testing show where respondents in a production survey may have difficulties or answer incorrectly and where revisions to the instrument may be required.

There has been an increasing amount of literature and reports recently related to cognitive testing of multilingual materials. Researchers have begun to focus on cognitive testing methods for use when testing in two languages, best practices for the management of multilingual pretesting, and methods for use in multilingual pretesting projects (Goerman and Caspar 2010; Goerman and Caspar 2010; Pan et al. 2009).

1.2. Behavior Coding

Behavior coding is the systematic coding of interviewer and respondent interactions in the field (Cannell et al. 1968; Sykes and Morton-Williams 1987; U.S. Census Bureau 2003). It identifies flawed questions by revealing question administration and response issues. Problems are detected by looking at rates of undesirable interviewer behavior, such as making changes to question wording, and undesirable respondent behavior, such as asking for clarification (suggesting that the question is not easy to understand without clarification). At the U.S. Census Bureau, we often use a rate of undesirable interviewer or respondent behavior that exceeds a particular threshold (e.g., 15 percent of cases) as an indication of a problem with a particular question (Fowler 1992; Landreth et al. 2006; Oksenberg et al. 1991).

1.3. Observational Study

Finally, the survey industry has recognized the value of observational studies that allow researchers to assess a survey instrument's performance by directly observing the interaction between the interviewer and respondent and noting problematic behaviors or circumstances (DeMaio 1983). Because behavior coding typically only captures verbal interactions via audio recording and leaves out nonverbal behaviors such as whether or not an interviewer shows a respondent a flashcard, an observational study was included as a part of the questionnaire development and evaluation described here. While the interviews were being recorded for the behavior coding study, researchers also observed and documented interviewer and respondent behavior related to several issues. The main goal of this observational study was to document flashcard use, but observers also noted several additional areas of interest including use of Spanish and/or English by interviewers and respondents prior to the start of the interview and nonverbal behaviors. For example, they made note of whether respondents answered questions by nodding or shaking their heads, which would not have been picked up on the audio recordings.

2. Methods

The pretesting cycle of the 2010 Census NRFU began with a field test in 2004 which contained a behavior-coding component. Based on results from the 2004 behavior coding research, cognitive testing with the self-administered paper form, and input from the U.S. Census Bureau's survey methodologists, the NRFU questions were modified for a subsequent field test in 2006. Just prior to the 2006 field test, the revised NRFU questionnaire was pretested via cognitive testing in both English and Spanish. Unfortunately, results of the cognitive testing were not complete in time to influence the questionnaire tested in the 2006 field test. As a part of the 2006 Census Test, a second behavior coding study was conducted with an observational study component. Finally, based on results from all of these studies, a revised questionnaire was developed and cognitively tested in 2007. Each of these steps is described in more detail below.

This article focuses on the cognitive testing, behavior coding and the observational studies because they were conducted by the authors. Additional field tests, evaluations, and studies were also conducted which led to improvements in the questionnaire and operational procedures, but those not conducted by the authors are not discussed here.

2.1. 2004 Behavior Coding

During the 2004 Census Test, a sample of interviews was tape recorded for behavior coding. We gathered 220 audio-taped interviews for assessment (119 English, 72 Spanish, and 29 mixed English and Spanish). Five bilingual telephone interviewers from a U.S. Census Bureau telephone center were trained by the research team in project-specific behavior-coding techniques. The research team included two monolingual English speakers and a bilingual English/Spanish speaker. Interviewer supervisors selected coders based on the supervisor's subjective assessment of their fluency in speaking and reading both English and Spanish, and on their reliability as interviewers. Audio-taped interviews were distributed among coders and each coder coded approximately 50 tapes. To assign

codes, coders listened to the audiotapes and followed along with a written guide that presented the questions in both languages. Coders made their judgments about the interviewer and response behavior based on the interactions only as they heard them on the audiotapes. They did not have access to the data that the interviewers entered in the CAPI instrument for each interview.

Behavior codes 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 the research of Oksenberg et al. (1991) and is attached in Appendix B. In addition to the codes themselves, when nonideal interactions occurred coders were instructed to transcribe or summarize the verbal interaction. These notes were then used for later qualitative analysis in which researchers analyzed the notes to identify problems and possible solutions.

In addition, the researchers assessed the reliability of the coders' work. Each coder coded the same eight interviews (four in each language) and their results were compared using the kappa statistic. The researchers noted that in the English-language interviews, the kappa scores ranged from .70 to .48 and reflected a good to fair level of agreement. However, reliability scores for the Spanish interviews reflected less reliable coding, ranging from .50 to .31, reflecting fair to poor agreement. To correct for this, researchers recoded a portion of the interviews based on verbatim notes made by the initial coders. A separate paper details possible causes for differences in reliability across the two languages (Goerman et al. 2008).

After the interviews were coded, the researchers conducted the analysis by first producing quantitative data on the percentage of times each behavior happened for each question (e.g., for Question 1, What percentage of the time did the interviewer read the question exactly as worded?; How often did she read it with a major change?; How often did the respondent answer without any problems?; How often did respondents ask for clarification?). We produced these data with both languages combined, but also separately for English and Spanish interviews. We then conducted a log linear regression to test for differences in interviewer and respondent behavior by language (described in more detail in Hunter and Landreth 2005 and Childs et al. 2007). For each question exhibiting interviewer or respondent behavior that we considered "poor" behavior (i.e., using a rule of thumb of 15% or more of undesirable behaviors), we conducted a qualitative analysis of all of the coder notes associated with the poor behavior. For example, for the first question in the 2006 behavior coding, interviewers made a major change to question wording in 26 percent of the instances. Therefore, we analyzed all of the notes of the cases where the interviewer made a major change to the question wording. During this analysis, we classified the verbatim notes into ad hoc categories and attempted to draw conclusions about the reason for the poor interviewing behavior based on how the interviewers rephrased the questions and how respondents reacted to the questions.

Coding English and Spanish cases allowed us to examine equivalency, or lack thereof, across the two language versions of the instrument. We could often identify areas where interviewers or respondents had more difficulty in one language than the other. Examples of types of findings generated by this study are included in the results section.

2.2. Cognitive Testing

Cognitive testing was conducted on both the English and Spanish versions of the 2006 NRFU instrument, but unfortunately, time, budget and staffing constraints made it impossible to conduct them jointly as Goerman and Caspar recommend (2010). Instead, staff from the U.S. Census Bureau conducted the English testing, and Development Associates, under contract with the U.S. Census Bureau, conducted the Spanish testing. Development Associates was provided with the Spanish version of the questionnaire, along with the protocol that was developed in English for the English-language interviews. They were asked to translate the protocol into Spanish, and adapt it as needed for the Spanish-language interviews.

2.2.1. English Testing

In 2005, U.S. Census Bureau staff conducted the first round of 14 cognitive interviews using a paper script of the 2006 NRFU questions (Hunter 2005). In the beginning of 2006, staff conducted a second round of 16 cognitive interviews using the 2006 NRFU instrument on the handheld computer (Childs et al. 2006). Respondents for the English testing were recruited to simulate the nonresponse population to the census. Respondents varied in age, race, and educational background and were interviewed in the Washington DC metropolitan area.

Both rounds of testing used a team of four interviewers led by the same lead researcher, but comprised different team members. As is standard practice at the U.S. Census Bureau, all cognitive interviewers had been trained in a 3-day course that focused on appropriate cognitive interviewing behaviors, such as nondirective and neutral probing. For each cognitive interview, one researcher was assigned the role of the survey interviewer, and the other was assigned the role of the cognitive interviewer. The survey interviewer administered the entire interview to the respondent while the cognitive interviewer observed. This allowed the cognitive interviewer to take detailed notes on respondent behaviors – both verbal and nonverbal – in order to plan probing questions for later in the interview.

The protocol for the cognitive interviews combined concurrent verbal reports with retrospective probes. Respondents were asked to think aloud while answering the questions, reporting any difficulty they were having answering or understanding any of the questions. Cognitive interviewers followed each set of questions with a series of probes. The interviews concluded with an additional set of retrospective probes, including a series of vignettes designed to explore respondents' understanding of key concepts from the questionnaire.

The cognitive interview protocol was semi-scripted, and took advantage of both scripted and emergent probing. Cognitive interviewers were free to follow up on interesting behaviors or responses as needed and to gather information to answer the scripted probes as they felt appropriate. Scripted probes included meaning-oriented probes, paraphrasing probes and expansive probes, in which interviewers asked about the respondent's personal situation to assess the match of his/her response to the survey question with his/her "true" situation.

The primary difference between the first and second rounds of English interviews was that the first was conducted with a paper script, and the second was conducted with the CAPI

instrument on the handheld computer. The use of both the paper script and the automated device allowed us to study the interactions with and without an automated intermediary.

For the analysis, each cognitive interviewer listened to audio-tape recordings of her own interviews to develop a detailed question-by-question interview summary with direct quotes from the respondent. After all summaries were completed, the researchers conducted a question-by-question analysis by comparing findings across cases. Findings from each round were predominately based on direct reports from respondents. Instances where the researchers made judgments based on their own experiences were carefully noted.

2.2.2. Spanish Testing

The Spanish script of the 2006 instrument was cognitively tested in two rounds, concurrently with the English, but this testing was done independently by different researchers. Two rounds of 15 Spanish interviews were conducted using paper script versions of the instrument (Beck 2006; Jones and Childs 2006). Respondents for the Spanish interviews were recruited to be monolingual Spanish speakers (or at least be more comfortable speaking in Spanish than in English). In the first round respondents were interviewed in southern California. In the second round respondents were interviewed in the Washington DC metropolitan area.

Each round was conducted by a single researcher, but the researcher differed by round. Both researchers were bilingual English/Spanish speakers. The researcher translated the protocol that was being used for the English interviews and administered it in Spanish. Therefore, the scripted probes were very similar, and the Spanish-speaking researchers were given the same instructions as were those who conducted the English interviews. The methodological differences between the English and Spanish interviews were as follows: (1) Only one researcher worked on each round of Spanish interviews (two researchers total); (2) A single researcher administered the survey interview questions and the cognitive interview probes in Spanish; (3) The Spanish language version was not tested on the handheld computer in either round of testing. This was not by design, but rather resulted from unforeseen technical issues with the computer.

The Spanish-speaking researchers were required to listen to their audiotapes and develop interview summaries comparable to those created in the English cases. The summaries were provided to the Census Bureau in Spanish with a translation into English. Each of the Spanish-speaking researchers also provided an overall assessment of their respective rounds of interviews. At this point, the English-speaking lead researcher conducted the global analysis of the Spanish cases in conjunction with the analysis of the English cases.

2.3. Observational Study

As previously mentioned, during the 2006 Census field test, an observational study was conducted in conjunction with gathering a sample of audiotapes for behavior coding. The researchers observed 99 eligible interviews, 65 in English and 34 in Spanish (Rappaport et al. 2006). Four bilingual English/Spanish researchers observed a total of 22 NRFU interviewers. Observers used a structured observation sheet that allowed them to record observations on several aspects of the interview, particularly flashcard and language use.

We did not assess the reliability of the observations – many of the observations required a subjective judgment by the researcher and each interview was only observed by one researcher, which made reliability coding impossible.

2.4. 2006 Behavior Coding

Unfortunately, only 72 of the 99 audiotapes that were recorded by the observational study researchers were usable for behavior coding; the rest were unusable for one of three reasons: (1) a failure to record respondents' consent on the audiotapes; (2) the taping of out-of-scope proxy interview; or (3) extremely poor audio quality of the recordings. The majority of the 72 usable cases were in English (54), but researchers also analyzed the 18 usable Spanish tapes (see Childs et al. 2007, for full results).

For the 2006 study, we used the same general method that we used for the 2004 behavior coding. The researchers conducting the study were two monolingual English speakers and one bilingual English/Spanish speaker. (The same two English-speaking researchers participated in both studies but a different bilingual researcher participated in each study). Five bilingual telephone interviewers from one of the U.S. Census Bureau telephone centers were trained by the researchers in project-specific behavior-coding techniques, and those interviewers served as the behavior coders. After training, staff turnover caused two interviewers to leave the project, so three coders completed the coding. Audio-taped interviews were divided among coders and each coder coded approximately 30 tapes. The coders' caseloads included duplicates of tapes used for reliability purposes. Those results are reported in Childs et al. (2007). Coders assigned codes and took notes, as described above under the 2004 behavior coding section, and analysis was conducted in the same way across the two studies.

2.5. Revised NRFU Cognitive Testing

Based on the results of the studies above, as well as results from the field tests themselves, a revised NRFU questionnaire was developed. A third and final round of cognitive testing was conducted in English only, with the revised, recommended NRFU script (Childs et al. 2007). Unfortunately, the revised questionnaire was not translated into Spanish to allow for cognitive testing prior to the deadline for the instrument to be finalized.

Six researchers conducted twenty-eight cognitive interviews with the revised NRFU questionnaire in the Washington DC metropolitan area. The same methods of testing and analysis were used this round as described with the earlier English cognitive interviews and a paper-and-pencil script was used.

The next section discusses the types of findings made possible by each of these pretesting methods in the case of the NRFU instrument.

3. General Findings

3.1. Cognitive Testing

Although the Spanish and the English cognitive testing were not done concurrently by the same researchers in a way that would provide two-way feedback during the testing, many findings were surprisingly similar. Questions that over-burdened interviewers and

respondents, problems with specific question concepts, and problems with the use of the flashcards were found across language versions of the survey.

3.1.1. Similar Findings Between Languages

Several of the questions in the 2006 version of the instrument were found to be too long for oral presentation in both languages. One example of this type of question is a question which asks respondents whether their unit is owned or rented. The question was scripted as follows:

Is this [house/apartment/mobile home]. . .
Owned by you or someone in this household with a mortgage or loan?
Owned by you or someone in this household free and clear?
Rented for cash rent?
Occupied without payment of cash rent?

¿Es [esta/este] [casa/apartamento/casa móvil]. . .
Propiedad suya o de alguien en este hogar con una hipoteca o préstamo?
Propiedad suya o de alguien en este hogar libre y sin deudas?
Alquilada por pago de dinero en efectivo?
Ocupada sin pago de dinero en efectivo?

While a lengthy question such as this one may work on a paper form, in the CAPI mode it requires a respondent to retain a great deal of information in working memory prior to formulating a response. Cognitive testing found that respondents often either asked for the question to be repeated or answered it incorrectly. These findings were consistent across the English and Spanish versions of the instrument (Childs et al. 2006; Hunter 2005; Jones and Childs 2006). As a result, we recommended shortening the question in order to improve interviewer ability to adhere to the script. A revised wording was tested in the English-only final round of cognitive testing:

Is this house owned by you or someone in this household?
Yes – Is it owned with a mortgage or owned free and clear?
No – Is it rented?

In the final round of testing, we found that respondents still had difficulty with this new question wording. The shorter length worked better, but respondents often focused on the "who" aspect of the question (e.g., do you own it or does someone else?; Childs et al. 2007). Based on this finding and because there was no time for another round of testing, the final question wording we recommended was based on a question used in another U.S. Census Bureau survey. It reads:

Do you or does someone in this household own this < house/apartment/mobile home > with a mortgage or loan (including home equity loans), own it free and clear, rent it or occupy it without having to pay rent?

Another finding that was similar across the two languages was that researchers commented that respondents seemed to need an introduction to the use of a flashcard (Childs et al. 2006; Hunter 2005; Jones and Childs 2006). In cognitive interviews, respondents in both languages expressed concern that they did not know if and when they

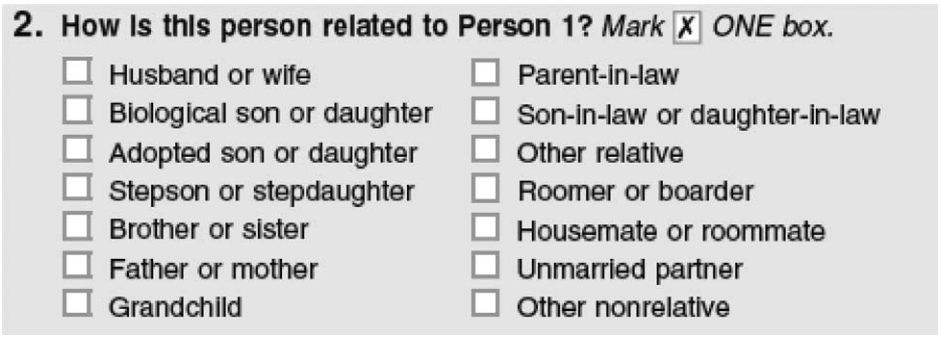


Fig. 1. Relationship Question on the Self-Administered Census Questionnaire

should read the information on a flashcard which gave supplemental instructions for "who to count" in their households. Interviewers handed respondents the card, then continued to read the scripted question without explicitly instructing them whether or not they should take time to read the information on the card before answering the question. Additionally, Jones and Childs (2006) noted that some respondents in particularly hard-to-enumerate populations, such as recent immigrants or those with low education, may have lower literacy levels and not be able to read the card (see also National Assessment of Adult Literacy 2006). These findings led to a recommendation to eliminate the use of flashcards whenever possible and to script the use of the cards in the interviewer text when it was necessary to use one, for example, saying "Using the guidelines on Card A, please tell me how many people are living or staying at this address."

The final example of a similar finding across the two languages deals with a question-level concept that was consistently interpreted in an unexpected way in both English and Spanish interviews. The "relationship question" is designed to record the relationships between the householder and all other residents of a household. During the testing cycle, the relationship question exhibited a series of problems. First, the CAPI instrument was a handheld computer with a small screen. This led to difficulty fitting all response categories from the paper form onto one screen. Figure 1 shows the layout of the relationship question on the self-administered paper census form.

a. Which one of these categories best describes how you are related to [NAME]?	**b. Which one of these categories best describes your relationship to [NAME]?**
Husband or wife	Roomer, boarder
Biological son or daughter	Housemate, roommate
Adopted son or daughter	Unmarried partner
Stepson or stepdaughter	Foster child or foster adult
Brother or sister	Other nonrelative
Father or mother	
Grandchild	
Parent-in-law	
Son-in-law or daughter-in-law	
Other relative	

Because the complete paper version of the question did not fit on one screen in the handheld computer instrument, the relationship question was modified to use a "branching" structure whereby respondents were first asked if two people were related:

Are you related to [NAME]?
Yes – Go to Question a
No – Go to Question b

Based on the answer to this question, respondents were skipped to either question a or b below:

The branched related-or-not-related question was found to be very problematic through cognitive testing in both English and Spanish. We found that respondents often did not categorize relationships in this prescribed manner, as "related" or "not related." For example, contrary to the U.S. Census Bureau's expectation, a proportion of respondents in both language groups classified spouses as "not related" to each other (Beck 2006; Hunter 2005). This proved to be problematic because after a respondent reported that his spouse was not "related" to him, he would be skipped to sub-question b, which did not include "wife" as an option. Similarly, both English and Spanish-speaking cognitive interview respondents disagreed with the Census Bureau's categorization of a number of relationships, including foster children, adopted children, and unmarried partners (Beck 2006; Hunter 2005; Jones and Childs 2006). The researchers expressed concern that going down the incorrect "related" or "not related" path might induce an interviewer or respondent to select an incorrect response option rather than going backwards in the instrument to find the more appropriate list of options. Anecdotal evidence suggests that interviewers try to avoid "backing up" in an instrument to prevent technical problems that often occur when backing up. This issue might be disproportionately problematic for Spanish-speaking respondents because we have evidence from other sources of testing that the Spanish translations currently being used for some of the non-relative categories are not working well with respondents (Caspar et al. 2007, Goerman et al. 2007).

This finding led to a recommendation not to branch the relationship question, but rather to ask the more general "How is NAME related to NAME?" and to use a flashcard in personal-visit interviews to help respondents who do not immediately choose a response from our list of options. Though a flashcard is not an ideal solution for respondents with low literacy, this particular use, to help generate a response when the respondent has difficulty, is one of the more straightforward uses of a flashcard. We also recommended that the interviewer be instructed to read the flashcard aloud when a respondent appears to have difficulty reading it.

3.1.2. Findings Unique to One Language

The Spanish cognitive testing study also uncovered translated terms that had conceptually inequivalent meanings to their English counterparts. An example of this is the term "residencia estacional," the translation used for "seasonal residence." In English, we found that this term was understood as intended, to mean a home that is used for particular seasons of the year, for example, a summer home. In Spanish, however, Spanish researchers found that the term "estacional" had a connotation of "stationary" or "parked," implying a permanence, that is opposite of the intended meaning (Jones and Childs 2006). In response to this finding, the researcher offered two different terms that might convey the

intended connotation better in Spanish – "temporal" or "de temporada" (which both mean "temporary" or "seasonal" in a way that adheres more closely to the English meaning).

3.2. Behavior Coding

We conducted two iterative rounds of behavior coding on the NRFU instrument. For both rounds of behavior coding, the English and Spanish language versions of the instrument were tested concurrently as a part of the same project with a team of researchers that included a bilingual researcher, so that results from one language could be directly compared to those from the other.

The behavior coding of the 2004 interviews showed statistically significant differences in good interviewer behavior across the two language versions of the survey – namely, interviewers administered the questions correctly more frequently when using the English than the Spanish version (Hunter and Landreth 2005). "Good" interviewer behavior was defined as asking questions exactly as worded, asking questions with minor changes, or correctly verifying information that had already been conveyed by the respondent. This finding held true for every question that we examined. This means that interviewers were better able to read the English questions as intended than they were the Spanish ones. We attributed this difference to three factors: (1) complex English wording which became even more complex through translation; (2) inexact translations; and (3) grammatical and/or spelling errors in the Spanish question wording on the instrument that was fielded. Errors in the Spanish instrument question wording were identified too late to be corrected before the field test.

Between the 2004 and 2006 field tests, some high level changes were made to the instrument, including changes to the sequence of questions and changing the structure of questions that asked about the household members' origin and race. These changes resulted from experimental field testing of those questions in a self-administered form that is not reported on here. Some additional changes were made to the English wording based on the 2004 testing, but not many changes were made to the Spanish question wording. Because of this, many of the same Spanish wording problems identified in 2004 were carried over to the 2006 instrument. As a result, the 2006 testing showed many of the same findings, and many of the same recommendations for the Spanish version were made after the 2006 field test.

In 2006, behavior coding again revealed statistically significant effects of language on overall interviewer behavior, but this time there were statistically significant effects of language on respondent and outcome behaviors as well (Childs et al. 2007). Questions in English were more often administered correctly than were those in Spanish. This trend was again evident for each of the questions that were examined. In English interviews, questions were asked in a good way 46 percent of the time, while they were asked in a similar way only 31 percent of the time in Spanish interviews. For respondent behavior, English questions yielded a rate of immediate adequate (or codable) response behavior 82 percent of the time, while Spanish questions yielded a rate of immediate adequate response behavior only 69 percent of the time. The fact that Spanish cases exhibited poorer interviewer and respondent behavior may be explained by a number of factors.

First, interviewer behavior may have been affected by the fact that the Spanish instrument was a translation and not an instrument initially developed in Spanish. This

may have caused it to sound less natural or conversational than the English version. Interviewers might have been trying to compensate for this by rewording some of the questions. Secondly, not all of the terms and questions in the Spanish instrument had been pretested prior to the fielding of the instruments to be sure that respondents would comprehend them as intended. This may have led interviewers to contextualize or alter question wording in places where they had found that questions did not "work" well with respondents in previous interviews. Another issue that may have affected interviewer behavior in Spanish is that there are different norms of politeness across cultures and it may not always have seemed appropriate to interviewers to launch into the scripted interview without making some small talk or framing questions in some way (see Rappaport et al. 2006, for a discussion on the "small talk" that occurred in each language prior to the survey).

Many of these same issues are likely to have had an impact on respondent behavior as well. For example, due to cultural conversational norms or difficulties with the translation, Spanish-speaking respondents might have felt that a discussion was warranted and they might have been less likely to give a brief response to the survey questions. Not surprisingly, we found that this was particularly the case in the Hispanic origin and Race questions. These questions have been shown to be particularly difficult for both English- and Spanish-speaking Hispanic respondents to answer in cognitive testing of different U.S. Census Bureau questionnaires in the past (see the example that follows about the Hispanic origin question and also see Caspar et al. 2007; and Goerman et al. 2007). To complicate the situation even more, Hispanic immigrant respondents with limited English proficiency often have lower educational levels than the average population in the U.S., and this may contribute to the need for greater discussion in answering the questions in Spanish.

An additional issue that may have had an impact on the coding of both the interviewer and respondent behavior is that the Spanish-speaking interviewers employed for the census test were not tested or certified as to their Spanish-language proficiency levels. In listening to some of the tapes, the researchers noticed that some Spanish-speaking interviewers had difficulty reading the Spanish questions aloud and had problems with Spanish pronunciation and grammar. It may have been difficult for coders to decide whether a question was read as intended by an interviewer when the interviewer had trouble pronouncing key terms in the question. Similarly, respondents may have had extra difficulty understanding and answering questions posed by interviewers with low levels of Spanish proficiency.

These behavior coding results make it clear that the Spanish versions of the questions did not perform as well as their English counterparts, which suggested to the researchers that they were in need of further revision and pretesting. Unfortunately, in this case study, the cognitive testing of the Spanish had not occurred in time to inform the wording used in the field tests.

3.2.1. Specific Example: Hispanic Origin Question

The U.S. Census Bureau's question on Hispanic origin has two objectives. The first is to identify each person as Hispanic or non-Hispanic. The second is, if the person is Hispanic, to identify their country of origin or ancestry. On the self-administered paper census form,

the Hispanic origin question has these two concepts embedded in the response categories. That question reads as follows:

Is Person 1 of Hispanic, Latino, or Spanish origin?
[ ] No, not of Spanish, Hispanic, or Latino origin
[ ] Yes, Mexican, Mexican American, or Chicano
[ ] Yes, Puerto Rican
[ ] Yes, Cuban
[ ] Yes, Another Hispanic, Latino, or Spanish origin, for example, Argentinean, Columbian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on. *Print origin.*_____

To adapt the question to an automated instrument in 2004, it was branched into a screener question with a follow-up question, as follows:

Are you of Hispanic, Latino or Spanish origin?
[ ] No
[ ] Yes —> Are you Mexican, Mexican American, or Chicano? Puerto Rican? Cuban? Another Spanish, Hispanic, or Latino origin? (For example, Argentinean, Columbian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on).

Behavior coding results showed that, surprisingly, Hispanics, and particularly Spanish-speaking Hispanics, did not always say "yes" in response to this question (Childs et al. 2007). In 2006, behavior coding showed high rates of respondents offering a nationality in response to this question rather than identifying themselves as "Hispanic" or saying "yes" (39% of Spanish-speaking respondents). Uncertainty as to how to answer the question may negatively impact data quality. For example, if a Hispanic respondent provides a nationality in response to the Hispanic origin question (instead of answering "yes"), it becomes problematic if the interviewer does not know whether the origin mentioned is a "Hispanic" origin. We witnessed an example of this during the 2006 behavior coding where a respondent answered "I'm Mexican" and the interviewer went on to verify with the respondent that she was therefore *not* of "Hispanic, Latino or Spanish origin" (Childs et al. 2007). Though this is a dramatic example, there are many Spanish-speaking countries that interviewers may not be familiar with or may not easily categorize as "Hispanic" countries, such as Uruguay, Bolivia or Ecuador. There are relatively fewer immigrants from those countries in the U.S. than from countries such as Mexico and they may not be as salient in the minds of interviewers without specialized training on the subject. In addition, there are countries that can cause confusion such as Brazil, which is a Latin American country but not a Spanish-speaking country, and thus not classified as "Hispanic" by the Census Bureau. Non-Hispanic respondents in the English language cognitive testing sometimes asked whether certain nationalities were considered Hispanic (e.g., Cuban or Italian; Hunter 2005). Since respondents are asked to report whether other household members are Hispanic, they may have difficulty and ask for clarification from interviewers. Thus cognitive testing and behavior coding indicate that the way this question is worded seems to place undue burden on both respondents and interviewers. Finally, a few respondents in both cognitive testing and behavior coding studies interpreted this question as citizenship question, which could cause privacy concerns that could even lead to nonresponse (Childs et al. 2007; Childs et al. 2007).

We hypothesized that when Hispanic respondents are speaking with an interviewer in Spanish, or are talking face-to-face with an interviewer in general, they may think that it should be obvious to the interviewer that they are "Hispanic." This context may lead them to interpret the question as a multiple choice question, asking whether they are (a) Hispanic, (b) Latino or (c) of Spanish origin. In fact, cognitive testing has also shown that many Hispanic respondents in both languages interpret the Hispanic origin question to be a multiple choice question rather than a yes/no question (Beck 2006; Childs et al. 2007; Jones and Childs 2006). Respondents often struggle to choose one of the three "options." This is in part because recent Spanish-speaking immigrants may not be familiar with the terms "Hispanic" and "Latino" because these are U.S. concepts that are not used in their home countries (Childs et al. 2007). In addition, when respondents hear the term "Spanish" they often think that the question is asking if they are "from Spain," which even leads some Spanish speakers to say "no" in response to the overall question (Childs et al. 2007). On the whole, we found that the way this question is worded is confusing for Hispanic respondents, particularly Spanish speakers.

We do not know how many respondents may answer "no" to this question incorrectly because they do not know that their country of origin is among those considered "Hispanic" or because they interpret the question to be asking whether they are "from Spain." Because the initial question requires only a yes or no response, there is some risk

Fig. 2. Recommendation for the Hispanic Origin Flashcard

that interviewers and respondents will not understand what is meant by "Hispanic, Latino or Spanish origin." For this reason, we recommended using a flashcard for this question. The flashcard presents the response categories as they appear in the self-administered paper form. This provides the respondents (and interviewers) with the same information provided to respondents in the self-response mode. Thus, we recommended that when answering the initial Hispanic origin question, respondents should see the list in Figure 2.

Unfortunately, this recommendation was made after the final round of cognitive testing (after seeing results without using the flashcard), and we did not have the opportunity to cognitively test the newly-worded flashcard. Despite that, the recommendation was adopted and this was used in the 2010 Census (see Appendix A).

3.3. Observational Study

Adding an observational study to our behavior coding research in 2006 offered invaluable information that would have been missed had we only analyzed interviewer and respondent interactions captured on audiotape. We gathered a great deal of information about one issue in particular, flashcard use.

The 2006 observational study provided the U.S. Census Bureau's most comprehensive and objective examination of census interviewer behavior with flashcards to date. Prior research has used interviewer debriefings to assess flashcard use in the field – which relies on a self-report. Cognitive testing has demonstrated the importance of flashcards as a visual aid (for a recent example, see Childs et al. 2007) but actual rates of flashcard usage in the field had not been previously studied to our knowledge.

In 2006, the NRFU interview employed three flashcards: (1) a flashcard that listed "Who to Count" to assist respondents in becoming aware of the Census Bureau's rules regarding who to count in a household for the census; (2) a "Relationship" flashcard that contained a list of possible relationships between the householder and other household residents; and (3) an "Ancestry" flashcard that contained an example list of origin or nationality categories. Interviewers were required to show all three flashcards to every respondent during the course of an interview.

The observers found that interviewers presented the "Who to Count" flashcard in only 25 percent of the 99 observed cases, the Relationship flashcard in only 28 percent and the Ancestry flashcard in 37 percent of cases (Rappaport et al. 2006). In 45 percent of the observed cases, the interviewer used at least one of the three flashcards. This indicates that interviewers were picking and choosing which flashcard to use in a given interview. In addition, this behavior differed by language. In English interviews, interviewers used the cards at rates of 28 percent, 25 percent, and 38 percent, respectively, whereas in Spanish, the rates were 17 percent, 33 percent, and 33 percent. Interestingly, interviewers used the "Who to Count" card somewhat less in Spanish interviews than in English ones. We judge this to be problematic since Spanish speakers in the U.S. are more likely to be immigrants and first generation immigrants more often live in mobile, complex households (Goerman 2005) for which creating a list of household residents is likely a more difficult task. Without the benefit of seeing all of the U.S. Census Bureau's rather complex rules, a respondent might be more likely to accidentally include someone who should not be included or omit a resident of his or her household when completing the

interview. The realization that interviewers were not consistently using this flashcard in the field led us to recommend changing the presentation of "Who to Count" rules from a flashcard to a series of shorter questions to be administered verbally, via automated instrument, to respondents. In this way, the questions could convey the same information without requiring the interviewer to show a card, or the respondent to read one. However, when the U.S. Census Bureau decided to use a paper-administered NRFU instrument, the "Who to Count" card was reinstated because of the difficulty of scripting a question-answer series to build a roster on a simple paper form. Thus, it was included in the 2010 Census and appears in Appendix A.

Because of documented difficulty interviewers have with using flashcards in a bound flashcard booklet and because we knew from the observational study that interviewers often chose not to use the flashcards at all, we recommended revising the format of the flashcards. It was anecdotally noted during observations of the field tests that interviewers did provide respondents with our legally required "confidentiality notice," which was printed on a single sheet of paper for the respondents to keep. Because we observed interviewers handing respondents the notice, but not using the flashcards, we decided to take advantage of their apparent willingness to hand respondents a sheet of paper. We therefore created a single "information sheet" for the respondents to keep that contains the confidentiality notice, as well as the flashcard "lists" for the each question that required a list. This new format was used in the 2010 Census (see Appendix A for the revised information sheet).

Demonstrating this alarmingly low rate of flashcard administration was a convincing argument to change the format of the flashcard. Actual flashcard use could only have been demonstrated through an observational study.

4. Conclusions

This case study shows the different types of results made possible through the application of different pretesting methods to the same bilingual survey instrument. While the timing and the sequencing of the studies were not ideal, examining the instrument's overall course of development allowed us to examine the types of findings made possible by the different pretesting methods and to recommend a more ideal sequence of testing for the future.

Cognitive testing took place in several rounds with the English and Spanish testing happening separately. The most interesting findings from those studies were the similarities between the results. Both English and Spanish speakers expressed difficulties with the administration of the "Who to Count" flashcard, as well as with the longer questions in the survey. In addition, the Spanish cognitive testing uncovered problems with conceptual equivalence between some of the Spanish and English terms used.

The behavior coding studies demonstrated how the survey was performing in the field in both languages directly compared to one another. In this case, the Spanish and English versions of the instrument were studied concurrently. The results pointed out problems with the Spanish instrument that were above and beyond the problems seen in the English survey and also showed where there was a lack of equivalency across the two language versions of the survey. Had the cognitive testing informed the wording in the survey

instrument that was fielded and behavior coded in 2004 and 2006, we might have seen fewer differences between language versions at the behavior coding stage. Finally, the behavior coding research brought to light problems in the U.S. Census Bureau's current hiring, assessment and monitoring procedures for non-English-language field interviewers.

The observational study went hand-in-hand with the behavior coding study and provided us with invaluable information about nonverbal aspects of the survey interview. From that study, we learned that interviewers were failing to show flashcards at alarming rates and we were able to implement a revision to the format of the flashcards to improve their administration.

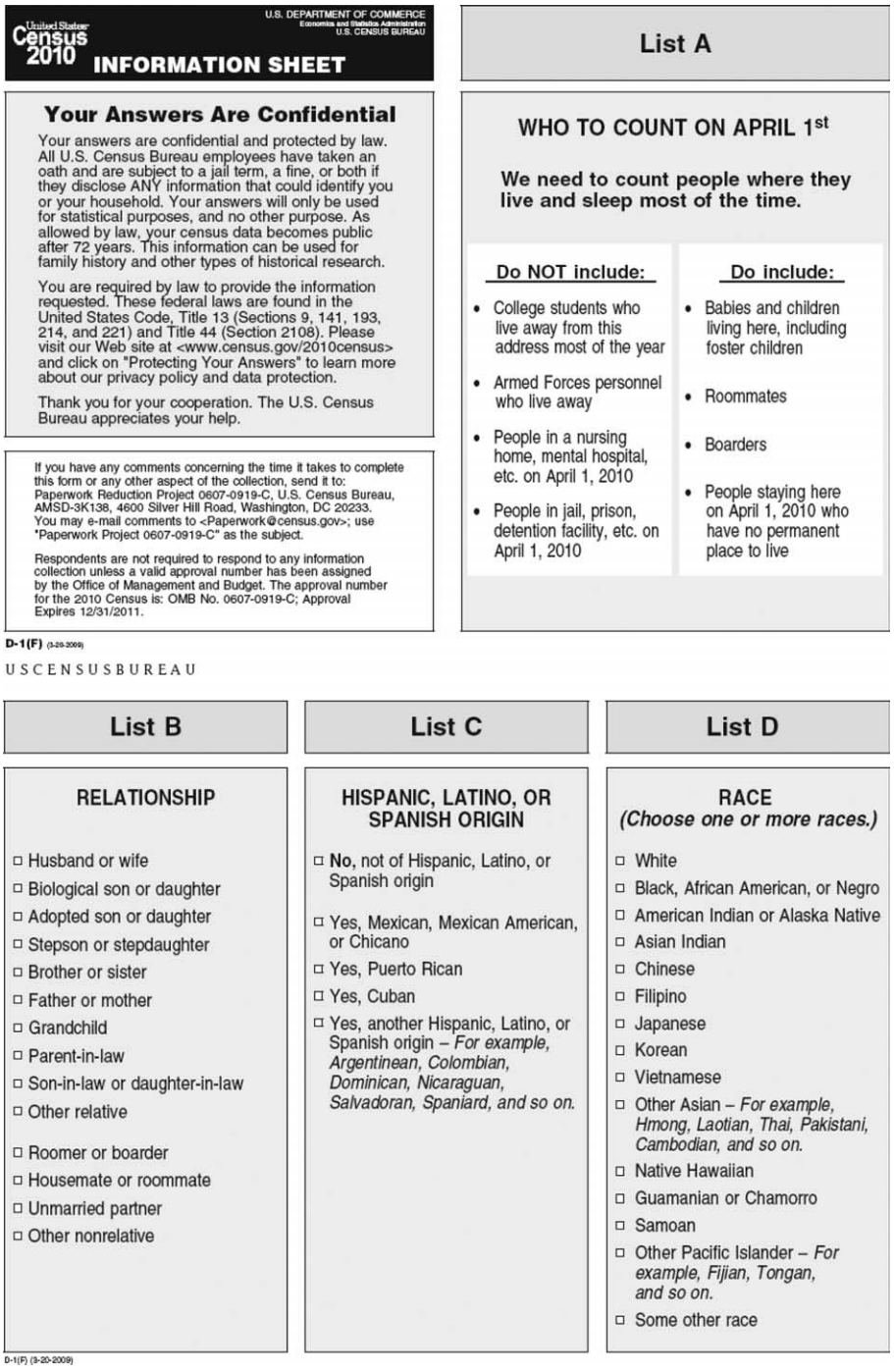
On the whole, each of the pretesting methods uncovered different types of issues and/or reinforced findings from other methods. They each provided information to assist researchers to improve the instrument in different ways. As a best practice, we recommend employing mixed methods of pretesting in the development of all survey instruments, but in particular, in the development of bilingual instruments. At the same time, we recommend a more in-depth examination of the ideal sequence of pretesting methods and we recommend better coordination across the methods than we were able to achieve in the development of this particular instrument.

4.1. Ideal Sequence for Multiple Pretesting Methods in the Development of a Bilingual Instrument

We recommend that prior to any field testing, translations be conducted or at least thoroughly reviewed using the committee approach (U.S. Census Bureau 2004). The next step should be concurrent iterative rounds of cognitive testing of both language versions (Goerman and Caspar 2007). Finally, behavior coding and an observational study should be conducted as a part of a field test to evaluate the question wording in both languages after it has been improved through cognitive testing. This recommended timeline for pretesting would allow for different types of improvements to be made to the questionnaires at each stage. The new wording could then be systematically tested at the next stage of development. Additionally, pretesting concurrently in both languages allows findings in each language to help improve the survey in the other language and to achieve better equivalence of meaning across language versions.

Despite the fact that we were not able to use the distinct pretesting methods in the ideal sequence in the development of the NRFU instruments, having used them all to study the same instrument has allowed us to have a well-rounded picture of how the survey would "work" in the field. We examined how the questionnaire performed in "real life" situations through the observational study as well as the behavior coding. Additionally, we looked into the minds of the respondents to see how they were interpreting the questions we were asking through the cognitive testing. Finally, this study enabled us to examine equivalency of meaning and interpretation across the source and translated versions of an instrument in each of these steps.

Appendix A. 2010 Census NRFU Revised Information Sheet



United States Census 2010

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

INFORMATION SHEET

Your Answers Are Confidential

Your answers are confidential and protected by law. All U.S. Census Bureau employees have taken an oath and are subject to a jail term, a fine, or both if they disclose ANY information that could identify you or your household. Your answers will only be used for statistical purposes, and no other purpose. As allowed by law, your census data becomes public after 72 years. This information can be used for family history and other types of historical research.

You are required by law to provide the information requested. These federal laws are found in the United States Code, Title 13 (Sections 9, 141, 193, 214, and 221) and Title 44 (Section 2108). Please visit our Web site at <www.census.gov/2010census> and click on "Protecting Your Answers" to learn more about our privacy policy and data protection.

Thank you for your cooperation. The U.S. Census Bureau appreciates your help.

If you have any comments concerning the time it takes to complete this form or any other aspect of the collection, send it to: Paperwork Reduction Project 0607-0919-C, U.S. Census Bureau, AMSD-3K138, 4600 Silver Hill Road, Washington, DC 20233. You may e-mail comments to <Paperwork@census.gov>; use "Paperwork Project 0607-0919-C" as the subject.

Respondents are not required to respond to any information collection unless a valid approval number has been assigned by the Office of Management and Budget. The approval number for the 2010 Census is: OMB No. 0607-0919-C; Approval Expires 12/31/2011.

D-1(F) (3-20-2009)

U S C E N S U S B U R E A U

List A

WHO TO COUNT ON APRIL 1st

We need to count people where they live and sleep most of the time.

Do NOT include:

- College students who live away from this address most of the year
- Armed Forces personnel who live away
- People in a nursing home, mental hospital, etc. on April 1, 2010
- People in jail, prison, detention facility, etc. on April 1, 2010

Do include:

- Babies and children living here, including foster children
- Roommates
- Boarders
- People staying here on April 1, 2010 who have no permanent place to live

List B

RELATIONSHIP

- ☐ Husband or wife
- ☐ Biological son or daughter
- ☐ Adopted son or daughter
- ☐ Stepson or stepdaughter
- ☐ Brother or sister
- ☐ Father or mother
- ☐ Grandchild
- ☐ Parent-in-law
- ☐ Son-in-law or daughter-in-law
- ☐ Other relative
- ☐ Roomer or boarder
- ☐ Housemate or roommate
- ☐ Unmarried partner
- ☐ Other nonrelative

List C

HISPANIC, LATINO, OR SPANISH ORIGIN

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican American, or Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *For example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

List D

RACE
(Choose one or more races.)

- ☐ White
- ☐ Black, African American, or Negro
- ☐ American Indian or Alaska Native
- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Other Asian – *For example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander – *For example, Fijian, Tongan, and so on.*
- ☐ Some other race

D-1(F) (3-20-2009)

Appendix B

Framework of Behavior Codes

Interviewer Behavior Codes

ES:	Exact Wording/Slight Change, interviewers read question exactly as worded or with slight change that did not affect question meaning
MC:	Major Change in Question Wording, interviewer changed the question in a way that could have changed the meaning of the question
V + :	Correct Verification, respondent provided information earlier that interviewer correctly verified and respondent accepts
V − :	Incorrect Verification, interviewer assumed or guessed at information not previously provided (even if correct) or misremembered information when verifying and respondent disagreed
IO:	Inaudible Interviewer/Other, interviewer exhibited some other behavior not captured under established codes
S:	Skipped question, interviewer failed to read a required question

Respondent Behavior Codes

AA:	Adequate Answer, respondent provided response that can easily be classified into one of the existing response options
IA:	Inadequate Answer, respondent provided a response that cannot easily be classified into one of the existing response options – often requiring interviewer to probe for more information
UA[2]:	Uncertain Answer, respondent expressed uncertainty about the response provided and may be unsure about the accuracy of the information
QA:	Qualified Answer, respondent placed conditions around their response (e.g., if you mean this, then answer is that)
CL[3]:	Clarification, respondent requested that a concept or question be stated more clearly
RR:	Question Re-Read, respondent asked interviewer to reread the question
DK:	Don't Know, respondent stated they do not have the information
R:	Refusal, respondent refused to provide a response
IO:	Inaudible Respondent/Other, respondent exhibited some other behavior not captured under established codes

A break-in code was also used to capture respondent behavior separately, and in addition to, the actual nature of the response/feedback.

Code BI: Break-In, respondent interrupted the reading of a question

[2] Codes UA and QA were combined into a single code for the 2006 coding because researchers determined that there was not a reliable distinction between the two codes.
[3] Codes CL and RR were also combined into a single code for the 2006 coding for the same reason.

5. References

Beatty, P. (2004). The Dynamics of Cognitive Interviewing. Methods for Testing and Evaluating Survey Questionnaires, S. Presser et al. (eds). New York: Wiley.

Beck, J. (2006). Cognitive Test of the 2006 Spanish NRFU: Round 1. Internal Report.

Cannell, C., Fowler, F., and Marquis, K. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on Reporting in Household Interviews. Vital and Health Statistics, Series 2 (26). Washington, DC: U.S. Government Printing Office.

Carrasco, L. (2003). The American Community Survey (ACS) en español: Using Cognitive Interviews to Test the Functional Equivalency of Questionnaire Translations. Statistical Research Division Study Series (Survey Methodology #2003-17). U.S. Census Bureau.

Caspar, R., Goerman, P., Sha, M., McAvinchey, G., and Quiroz, R. (2007). Census Bilingual Questionnaire Research Final Round 1 Report. U.S. Census Bureau. Statistical Research Division Report Series (Survey Methodology # 2008-1). U.S. Census Bureau.

Childs, J.H. and Landreth, A. (2006). Analyzing Interviewer/Respondent Interactions While Using a Mobile Computer-Assisted Personal Interview Device. Field Methods, 18(3), 335–351.

Childs, J.H., Carter, III., G.R., Norris, D., Hanaoka, K., and Schwede, L. (2007). Cognitive Test of the NRFU Round 3: Revised Questions. Statistical Research Division Report Series (Survey Methodology # 2007-9). U.S. Census Bureau.

Childs, J.H., Gerber, E.R., Carter, G., and Beck, J. (2006). Cognitive Test of the 2006 NRFU: Round 2. Statistical Research Division Study Series (Survey Methodology # 2006-5). U.S. Census Bureau.

Childs, J.H., Landreth, A., Goerman, P., Norris, D., and Dajani, A. (2007). Behavior Coding Analysis Report: Evaluating the English and the Spanish Versions of the Non-Response Follow-Up (NRFU) for the 2006 Census Test. Statistical Research Division Study Series (Survey Methodology # 2007-16). U.S. Census Bureau.

DeMaio, T.J. (ed.) (1983). Approaches to Developing Questionnaires. Statistical Policy Working Paper 10. Washington, DC: Office of Management and Budget.

DeMaio, T.J. and Rothgeb, J.M. (1996). Cognitive Interviewing Techniques: In the Lab and in the Field. Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass, 177–196.

Dillman, D.A. and Christian, L.M. (2005). Survey Mode as a Source of Instability in Responses Across Surveys. Field Methods, 17, 30–52.

Fowler, F. (1992). How Unclear Terms Affect Survey Data. Public Opinion Quarterly, 56, 218–231.

Goerman, P. (2005). Making Ends Meet: The Complex Household as a Temporary Survival Strategy Among New Latino Immigrants to Virginia. Complex Ethnic Households in America, L. Schwede, R.L. Blumberg, and A.Y. Chan (eds). Lanham, MD: Rowman & Littlefield Publishers, Inc., 149–180.

Goerman, P. (2006). Adapting Cognitive Interview Techniques for Use in Pretesting Spanish Language Survey Instruments. Statistical Research Division Research Report Series, (Survey Methodology #2006-3). U.S. Census Bureau.

Goerman, P.L. and Caspar, R. (2010). A Preferred Approach for the Cognitive Testing of Translated Materials: Testing the Source Version as a Basis for Comparison. International Journal of Social Research Methodology, 13, 303–316. First published on November 16, 2009 (iFirst).

Goerman, P.L. and Caspar, R. (2010). Managing the Cognitive Pretesting of Multilingual Survey Instruments: A Case Study Based on the Pretesting of the U.S. Census Bureau Bilingual Spanish/English Questionnaire. Survey Methods in Multinational, Multiregional, and Multicultural Contexts, J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L.E. Lyberg, P.P. Mohler, B. Pennell, and T. Smith (eds). New York, NY: John Wiley & Sons, Inc.

Goerman, P.L., Caspar, R., Sha, M., McAvinchey, G., and Quiroz, R. (2007). Census Bilingual Questionnaire Research Final Round 2 Report. Statistical Research Division Report Series (Survey Methodology # 2007-27). U.S. Census Bureau.

Goerman, P.L., Childs, J.H., and Clifton, M. (2008). Explaining Differences in Inter-coder Reliability between English and Spanish Language Behavior Coding Research. Paper presented at the American Association for Public Opinion Research conference, May 14–18, New Orleans, Louisiana. 2008 JSM Proceedings, Statistical Computing Section [CD-ROM], 4156-4163. Alexandria, VA: American Statistical Association.

Harkness, J.A. (2004). Problems in Establishing Conceptually Equivalent Health Definitions Across Multiple Cultural Groups. Paper presented at the Eighth Conference on Health Survey Research Methods. Cohen, S.B. and Lepkowski, J.M. (eds.) Hyattsville, MD: National Center for Health Statistics.

Harkness, J.A., Van de Vijver, F.J.R., and Mohler, P.Ph. (eds) (2003). Cross-Cultural Survey Methods. Hoboken, New Jersey: John Wiley & Sons.

Hunter, J. (2005). Cognitive Test of the 2006 NRFU: Round 1. Statistical Research Division Study Series Report, (Survey Methodology #2005-07). U.S. Census Bureau.

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. Statistical Research Division Study Series Report, (Survey Methodology #2006-07). U.S. Census Bureau.

Jones, J.C. and Childs, J.H. (2006). Final Report on Round 2 of the Spanish 2006 NRFU Cognitive Test. Internal Report for the U.S. Census Bureau.

Landreth, A.D., Krejsa, E.A., and Karl, L. (2006). Behavior Coding Analysis Report: Evaluating the Coverage Research Follow-Up (CRFU) Survey for the 2004 Census Test Administered using Telephone and Personal Visit Survey Modes. Statistical Research Division Study Series Report, (Survey Methodology #2006-01). U.S. Census Bureau.

Martin, E., Childs, J.H., DeMaio, T., Hill, J., Reiser, C., Gerber, G., Styles, K., and Dillman, D.A. (2007). Guidelines for Designing Questionnaires for Administration in Different Modes. Washington, DC: U.S. Census Bureau, 20233.

National Assessment of Adult Literacy. (2006). A First Look at the Literacy of America's Adults in the 21st Century. National Center for Education Statistics, Institute of Education Science. (NCES 2006-470). U.S. Department of Education.

Nicholls, W.L. and de Leeuw, E. (1996). Factors in Acceptance of Computer-assisted Interviewing Methods: A Conceptual and Historic Review. Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA, 758–763.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New Strategies for Pretesting Survey Question. Journal of Official Statistics, 7, 349–394.

Pan, Y. (2004). Cognitive Interviews in Languages Other Than English: Methodological and Research Issues. 2004 American Statistical Association Proceedings of the Joint Statistical Meetings. Phoenix, AZ. May 13–16.

Pan, Y., Sha, M., Park, H., and Schoua-Glusberg, A. (2009) 2010 Census Language Program: Pretesting of Census 2010 Questionnaire in Five Languages. U.S. Census Bureau Statistical Research Division Research Report Series (Survey Methodology – #2009-1), U.S. Census Bureau.

Potaka, L. and Cochrane, S. (2004). Developing Bilingual Questionnaires: Experiences from New Zealand in the Development of the 2001 Maori Language Survey. Journal of Official Statistics, 20, 289–300.

Rappaport, M., Davis, D., and Allen, S. (2006). Final Report on an Observational Study of Census Non-response Follow-up Interviews during the 2006 Census Test in Travis County, Texas Spring and Summer. Report submitted by Development Associates to the Census Bureau, December 4.

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

Willis, G. (2004). Overview of Methods for Developing Equivalent Measures Across Multiple Cultural Groups. Paper presented at the Eighth Conference on Health Survey Research Methods. Cohen, S.B. and Lepkowski, J.M. (eds.) Hyattsville, MD: National Center for Health Statistics.

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

U.S. Census Bureau (2003). Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses. U.S. Department of Commerce. Washington, DC: Author.

U.S. Census Bureau (2004). Census Bureau Guideline: Language Translation of Data Collection Instruments and Supporting Materials. U.S. Department of Commerce. Washington, DC: Author.

Received November 2008
Revised December 2009