RESEARCH REPORT SERIES
(Survey Methodology #2008-07)

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

Jennifer Childs
Patricia Goerman

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

Report Issued: September 30, 2008

Multilingual Questionnaire Evaluation and Development through Mixed Pretesting Methods: The Case of the U.S. Census Nonresponse Followup Instrument[1]

Jennifer Childs and Patricia Goerman

U.S. Census Bureau

September 26, 2008

[1] 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 paper was written prior to a decision not to use a CAPI instrument for Nonresponse Followup in the 2010 Census as had originally been planned. This paper discusses research and plans as they were prior to that decision. Many of the findings will inform the development of a paper Nonresponse Followup form.

Abstract

The objective of the research described in this paper was to develop and improve a Nonresponse Followup (NRFU) instrument for the 2010 U.S. Census. This research is unique in that multiple pretesting methods were used in the development of a single instrument in two different languages: English and Spanish.

The NRFU instrument was originally developed and tested as a computer-assisted personal interview (CAPI) instrument. The U.S. Census Bureau's NRFU operation is critical to gathering data from both English- and Spanish-speaking households who do not return their self-administered census forms. The instrument was developed using a series of pretesting methods, including usability and cognitive testing, behavior coding, and an observational study of the administration of test versions of both the English and Spanish instruments. Though many of the questions had already been tested in the context of a self-administered paper form, this was the first time pretesting was conducted using the interviewer-administered questions as they were scripted to be read from a CAPI instrument.

This paper discusses overarching results of three rounds of English cognitive testing, two rounds of Spanish cognitive testing, two rounds of usability testing in both languages, two rounds of behavior coding of the instrument in both languages, and an observational study of the administration of the NRFU interview 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 paper 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.

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

Pretesting of 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 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 (2003) requires that survey questions be pretested and shown to "work" prior to being fielded. Both the Census Bureau standards and guidelines recommend pretesting questions in the languages in which they will be administered.

The objective of this paper 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) 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.

Background

As a part of the decennial census operations, the Census Bureau mails out forms to most housing units in the country. The Census Bureau attempts to send an interviewer to every known 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, self-administered paper census forms and the CAPI NRFU instrument have been developed in both English and Spanish. The development of the bilingual

CAPI instrument is the focus of this paper.[2] In the development cycle, the self-administered census questionnaire that will be mailed to respondents was the first thing to be developed. 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. Those questions are the focus of this pretesting effort.

The 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 also includes flashcards created to assist respondents in answering particularly long or complex questions. Flashcards are used to present instructions on "who to list" in the household, the relationship between the householder and other residents, and the various origin and race response categories included in the survey instrument.

Pretesting Timeline

The NRFU instrument was developed first in English, and then translated into Spanish. While this is not the ideal way to develop a multilingual survey instrument,[3] both cost and staffing resources influenced the decision to develop the English instrument first. Pretesting began after the questionnaire had been developed in English, translated into Spanish and programmed into the CAPI instrument in both languages.

The pretesting cycle of the NRFU began with a field test in 2004. During that field test, a sample of interviews was tape recorded for behavior coding. We gathered 220 audio-taped interviews (119 English, 72 Spanish, and 29 mixed English and Spanish). These results are documented fully in Hunter and Landreth (2005). Based in part on results from the 2004 behavior coding research, separate cognitive testing with the self-administered paper form, and input from the Census Bureau's survey methodologists, the NRFU questions were modified between the 2004 field test and the second field test which occurred in 2006.

[2] The 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 will now be collected via an interviewer administered paper-and-pencil instrument. Nevertheless, this paper discusses lessons learned from the CAPI instrument development process, which will inform future Census Bureau initiatives.

[3] See Potaka and Cochrane, (2004) for discussion of ideal methods for developing bilingual instruments.

Among the changes was a shift from a "person-based" to a "topic-based" format. A "person-based" approach means that a series of questions is initially asked about the first person living in the household and then about each subsequent person in turn. For example, a respondent is first asked for his or her name, age, date of birth and race. Then, the interviewer asks for a second person's name, age, date of birth and race. A "topic-based" approach means that data regarding a single topic are gathered for everyone in the household before moving on to the next topic in the survey. In this method, the interviewer would ask first for the name of each person in the household, then for each person's age, then for each person's race and so on. The self-administered paper census form employs a person-based format, with a column for each individual person in a household. Because the 2004 NRFU instrument was first developed based on the content of the paper form, it was initially created in the same person-based format.

Prior to the field test in 2006, a revised NRFU instrument was pretested via usability and cognitive testing. As a part of the 2006 Census Test, a second behavior coding study was conducted and this time an observational study was included as well. Each of these steps is described in greater detail below.

Two rounds of usability testing were conducted with two early versions of the 2006 NRFU instruments in the summer of 2005. The first round of usability testing had six participants: four English speakers and two Spanish speakers (Olmsted, Hourcade and Abdalla, 2005). The second round had five participants: four English speakers and one Spanish speaker (Olmsted and Hourcade, 2005). Results from this usability study influenced the visual layout of the 2006 instrument that was field tested.

At about the same time as the usability study, two rounds of cognitive testing were conducted on the 2006 NRFU wording in English. The first 14 interviews were conducted using a paper script in the summer of 2005 (Hunter, 2005), and in the beginning of 2006, the second round, consisting of 16 interviews, was conducted using the 2006 NRFU instrument as it was programmed for administration via handheld computer (Childs, Gerber, Carter and Beck, 2006).

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).

The cognitive testing studies were not conducted in time to influence the 2006 NRFU instrument wording prior to the field test. However, those findings were compiled with findings from the behavior coding and observational studies conducted in 2006 to generate recommendations for 2008 and beyond.

During the 2006 Census 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, Davis and Allen, 2006). Unfortunately, only 72 of the audiotapes that were recorded were usable for behavior coding; the rest were unusable for one of three reasons: a failure to record respondents' consent on the audiotapes; the inadvertent taping of proxy interviews, or the extremely poor audio quality of the recordings. The majority of the 72 usable cases were in English (54), but analysis was also conducted on the 18 usable Spanish tapes (Childs, Landreth, Goerman, Norris and Dajani, 2007).

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 2008 NRFU script (Childs, Carter, Norris, Hanaoka, and Schwede, 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 for 2008.

Methods

In this section, each pretesting method is described very generally. More detailed study-specific methods can be found in the individual study reports.

Usability Testing

The goal of usability testing is to improve the usability of a product so that "the *people who use the product* can do so *quickly and easily*" (p. 4, Dumas and Reddish, 1999). In usability testing,

the participant often plays the part of an interviewer and is given a limited amount of training on how to administer the instrument. The participant is then asked to administer the instrument to respondents as an "interviewer" in order to gauge the ease with which the participant/ interviewer can "use" the instrument. The "respondents" are generally played by the researchers, using prearranged respondent scripts. The goal is to evaluate whether the instrument is "usable," i.e., intuitive enough for someone with limited training to be able to navigate without many problems. Usability testing was conducted by the Census Bureau on the 2006 English and Spanish NRFU instruments.

Cognitive Testing

Cognitive testing is a method by which participants are administered a survey, usually in a lab setting, and are asked concurrent or retrospective probes about their thought processes while answering the questions. Results from cognitive testing show us where respondents in a production survey may have difficulties or answer incorrectly and where revisions to the instrument may be required. See Willis (2005) for a detailed explanation of cognitive testing as a pretest method. Goerman and Caspar (2007) discuss cognitive testing methods for use when testing in more than one language. Cognitive testing was conducted on both the English and Spanish versions of the 2006 NRFU script, but unfortunately, time, budget and staffing constraints made it impossible to conduct them jointly as Goerman and Casper recommend.

Behavior Coding

Behavior coding is the systematic coding of interviewer and respondent interactions in the field (Cannell, Fowler, & Marquis, 1968). It identifies flawed questions by revealing 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). Undesirable interviewer or respondent behavior that exceeds 15 percent of cases is deemed an indication of a problem with a particular question (Oksenberg, Cannell, & Kalton, 1991; Fowler, 1992).

As previously mentioned, the two behavior coding studies were conducted based on tape recordings of field interviews that were conducted as a part of the 2004 and 2006 tests of the NRFU instrument. Census Bureau telephone interviewers were trained on how to apply behavior codes that described interviewer and respondent behavior while listening to the audiotaped interviews and the coders were trained to take detailed notes whenever a non-ideal interaction between an interviewer and respondent occurred. Qualitative analysis of the coders' notes allowed us to see exactly where problems occurred and to hypothesize about how these problems might be solved. Looking at behavior coding data from the 2004 and 2006 tests, we focused on four major behaviors: 1) interviewer behavior (i.e., whether interviewers administered questions exactly as scripted in the instrument); 2) respondent behavior (i.e., whether or not the respondent provided a codable answer as his or her first response); 3) whether or not the respondent interrupted the interviewer during the reading of the question, which we call a "break-in"; and 4) the final outcome, (i.e., whether the interviewer and respondent arrived at a codable response by the end of the interaction). 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.

Observational Study

Because behavior coding only captures verbal interactions and leaves out gestures, facial expressions and actions such as whether or not an interviewer shows a respondent a flashcard, an observational study was included as a part of the 2006 Census Test of the NRFU instrument. While the interviews were being recorded for the behavior coding study, researchers also observed and documented interviewer and respondent behavior related to several key issues. The main goals of the observational study were to document flashcard use, language use, and other non-verbal behaviors (such as answering questions by nodding or shaking the head) that would not have been picked up on the audio recordings.

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

General Findings

Usability Testing

The usability studies on the NRFU instrument provided an early glimpse of many issues that we would later explore through cognitive testing. Olmsted et al. (2005) noted that reading topic-based questions over and over in full for each household member seemed repetitive and burdensome for the interviewer. This finding was later replicated through the cognitive interview and behavior coding studies.

Olmsted and Hourcade (2005) also documented difficulties in working with flashcards given that their use was not scripted in the instrument itself. For example, some survey instruments include a statement such as "Please look at Card A while I read the next question" as a part of the question wording to be read aloud to respondents. The 2004 and 2006 versions of the NRFU instrument included instructions to the interviewer to show the flashcard to the respondent but did not include any text to read to the respondent to introduce the flashcard use. This often caused a problem because there was no scripted pause to allow for respondents to actually read the flashcard. This problem was also observed during the cognitive testing and observational studies.

Finally, the usability research done by Olmsted et al. (2005) led researchers to conclude that the Spanish translation of the instrument sounded unnatural. Overall, the researchers felt that the Spanish question wording sounded "correct but overly literal." Additionally, Olmsted and Hourcade (2005) anticipated possible difficulties for recent immigrants who wish to report dates of birth in the sequence of: day, month and then year instead of the American format of month, day and year. Since the usability testing did not focus in-depth on question wording, they recommended cognitive testing of the Spanish language questionnaire to identify specific problems and possible improved wording prior to fielding the instrument.

Cognitive Testing

Although the Spanish and English cognitive testing was not done concurrently by the same researchers in a way that would provide two-way feedback during the testing, many findings were surprisingly similar. Issues such as 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.

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 well on a paper form, in 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. 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, Carter, et al., 2007). Based on this finding, the final recommended question was based on a question used in another 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?**

As a result of both the English and Spanish cognitive testing studies, researchers commented that respondents seemed to need an introduction to the use of a flashcard, including actual scripted time for the respondent to read the information on the card prior to being asked a question (Childs, Gerber, et al., 2006; Hunter, 2005; and Jones and Childs, 2006). In cognitive interviews, respondents in both languages expressed concern that they did not know if and when they should read the information on the card. Additionally, Jones and Childs (2006) noted that some respondents in 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). This finding led to recommendations of eliminating the flashcards whenever possible and scripting the use of the cards in the interviewer text when it was necessary to use one.

The Spanish research 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, like a summer home. In Spanish, however, 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).

In addition to these Spanish-specific, or translation-related findings, other misunderstandings of larger question concepts were uncovered in the separate cognitive testing studies. The next section presents an in-depth analysis of a specific example, based on data from English and Spanish cognitive interviews.

Example: The Relationship Question

One of the questions in the NRFU instrument is designed to record the relationships between the householder and all other residents of a household. We call this the "relationship question." During the testing cycle, the relationship question exhibited problems on several fronts. First, the CAPI instrument for which the question was originally designed was a handheld computer with a small screen. This led to difficulty in fitting all response categories from the paper form into one screen shot. See Figure 1 for the layout of the relationship question on the self-administered paper census form.

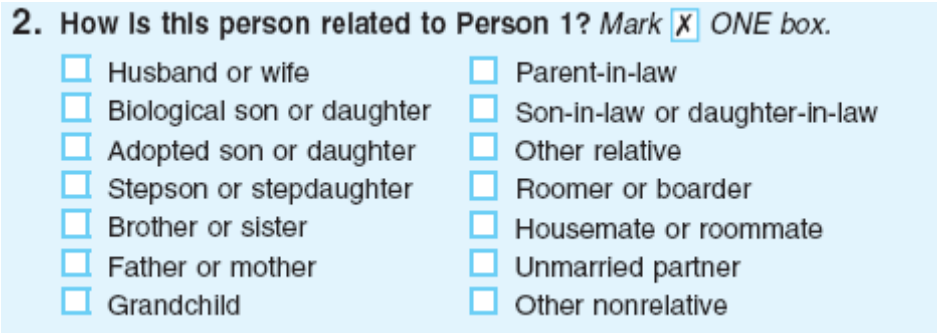

Figure 1. Relationship Question on the Self-Administered Census Questionnaire

Because the complete 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 a
No – Go to b

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

a. Which one of these categories best describes how you are related to [NAME]?

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

b. Which one of these categories best describes your relationship to [NAME]?

Roomer, boarder
Housemate, roommate
Unmarried partner
Foster child or foster adult
Other nonrelative

The branched related-or-not-related questions were found to be very problematic through cognitive testing in both English and Spanish. We found that respondents often do not categorize relationships in this prescribed manner, as "related" or "not related." For example, contrary to the 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 since 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 & Childs, 2006). The researchers expressed concern that going down the incorrect "related" or "not related" path might induce an interviewer to select an incorrect response option rather than going backwards in the instrument to find the more appropriate list of options. This issue might be disproportionately problematic for Spanish-speaking respondents since we have evidence 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).

These findings led to a recommendation not to branch the relationship question, but rather to ask the more general "How is NAME related to NAME?" or "¿Cómo está NAME relacionado(a) con 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 recommended that the interviewer be instructed to read the flashcard aloud when a respondent appears to have difficulty reading it.

Behavior Coding

Behavior coding was the third testing method that we applied to the NRFU instrument in this development cycle. We conducted two iterative rounds of behavior coding on different versions of the instrument. When using this method, the English and Spanish language versions of the instrument were tested concurrently as a part of the same project, so that results from one language could inform recommendations for the other.

The behavior coding of the 2004 interviews showed differences in good interviewer behavior across the two language versions of the survey--namely, interviewers read the question text exactly as scripted more frequently when using the English than the Spanish version (Hunter & Landreth, 2005). 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 these differences to three factors: complex English wording which became even more complex through translation; inexact translations; and errors in the Spanish translated instrument that was fielded.

Between the 2004 and 2006 field tests, some high level changes were made to the instrument, such as switching from a person-based to a topic-based format and changing the structure of questions that asked about the household members' origin and race.[4] Some changes were made to the English wording based on the 2004 testing, but not many changes were made to Spanish question wording. Because of this, many of the same Spanish wording problems identified in

[4] The change to the origin and race questions resulted from separate testing of the same questions in a self-administered form.

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 Spanish recommendations were made after the 2006 field test.

In 2006, behavior coding again revealed significant effects of language on overall interviewer behavior, but this time there were also significant effects of language on respondent and outcome behaviors as well (Childs, Landreth, 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[5] 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 adequate (or codable) response behavior 82 percent of the time, while Spanish questions yielded a rate of adequate response behavior only 69 percent of the time. Final outcome behavior was similar with an adequate final outcome in 89 percent of English interactions but in only 79 percent of Spanish interactions. The fact that Spanish cases exhibited poorer interviewer and respondent behavior and outcomes may be explained by a number of factors.

First, interviewer behavior may have been affected by the fact that the Spanish instrument is a translation and not an instrument initially developed in Spanish. This may cause it to sound less natural or conversational than the English version. Interviewers might be 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 properly 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 seem 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).

[5] "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.

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 Census Bureau instruments 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 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 Spanish 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 version 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.

Example: Hispanic Origin

The 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 to identify the person's country of origin (or

ancestry). On the self-administered census form, the Hispanic origin question has two concepts embedded in the response categories: whether or not someone is of Hispanic origin and his or her specific national origin. The 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 followup 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 have shown that, surprisingly, Hispanics, and particularly Spanish-speaking Hispanics, do not always say "yes" in response to this question (Childs, Landreth, 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). 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, Landreth, et al., 2007; Jones & 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" since these are U.S. concepts that are not used in their home countries (Childs, Landreth, 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, Landreth, et al., 2007).

Interpreting this as a multiple choice question may at best cause unnecessary respondent burden, and at worst could negatively impact data quality. 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, Landreth, et al., 2007). Though this is a dramatic example, there are many Spanish-speaking countries that field 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. In addition there are examples 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. We found 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 non-response (Childs, Carter, et al., 2007; Childs, Landreth, 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 is a yes/no only, there is some risk 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, it was recommended that when answering the initial Hispanic origin question, respondents should see a list on their information sheet that looks like this:

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*

Unfortunately, this recommendation was made after the final round of cognitive testing, and we did not have a chance to cognitively test the newly-worded flashcard prior to the deadline for finalizing the 2008 instrument.

Observational Study

Adding an observational study to our behavior coding and field test research in 2006 offered invaluable information that would have been missed had we only been able to analyze interviewer and respondent interaction on audiotape after the fact. We gathered a great deal of information about two issues in particular: flashcard use and language use (e.g., whether bilingual interviewers or respondents switched back and forth between Spanish and English through the course of an interview). Though language use could be examined through the audiotapes, it was not one of the analytical questions in the behavior coding study and therefore it was not analyzed there.

Flashcard Use

The 2006 field test observational study provided the Census Bureau's most comprehensive examination of interviewer behavior with flashcards to date. In 2006, the NRFU interview employed three supplementary 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. The interviewers were required to show all three flashcards to all respondents during the course of the interviews. A total of 99 interviews were observed as a part of the observation study.

The observers found that the "Who to Count" flashcard was presented in only 25 percent of 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, at least one of the three flashcards was used. 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, the cards were used 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, the "Who to Count" card was used somewhat less in Spanish interviews than in English ones. We judged 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 might be a more difficult task. Without the benefit of seeing all of the Census Bureau's rather complex residence 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 Census Bureau could convey the same information without requiring the interviewer to show a card, or the respondent to read one.

In revising the instrument for 2008, there was still a need for flashcards for the relationship and ancestry (more specifically, origin and race) questions because of their lengthy response sets and the difficulty it would pose for interviewers to read the entire response sets to all respondents. 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 revised the format of the flashcards. It was 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 Relationship, Hispanic origin, and Race questions. This new format will be used in the 2010 Census. See Appendix A for an example of a draft of this new information sheet.

Interviewer Language Use and Proficiency

Finally, the observational study provided a unique view of English and Spanish language interviews in predominantly Hispanic neighborhoods. Most of the interviews observed were conducted entirely in English (62%), and about a third (30%) were conducted in Spanish (Rappaport et al., 2006). The language of the interview changed between English and Spanish during the interview in only eight percent of interviews observed. In four (of 99) cases, language changed more than once during the interview. Rappaport et al. noted that the interviewer often was the person who initiated the language switch. Specific Spanish language problems noted by Rappaport et al. included the interviewer not being fully fluent in Spanish, stumbling over pronunciations, and substituting English words in Spanish interviews when the Spanish word was unfamiliar to them. This information allowed us to make recommendations as to problems that can occur when field interviewers are not screened for language proficiency prior to conducting interviews in a non-English language in the field. Additionally, this information can be used to inform future behavior coding trainings, as it can be difficult for coders to decide how to code interactions in which a conversation shifts from one language to another.

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 sequencing of the different studies was 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.

The pretesting of the NRFU instrument started with usability testing in both English and Spanish. This study, while providing important usability data not covered in this paper, also provided an inkling of the kinds of cognitive findings we could expect in later testing. (See Olmsted et al., 2005 and Olmsted and Hourcade, 2005, for complete information on the usability results.)

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 "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. 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 in many cases. 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 this stage. In addition, the behavior coding research brought to light problems in the 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 non-verbal and unrecorded aspects of the survey interview. From that study, we learned that interviewers were failing to show flashcards at alarming rates. We also saw evidence of interviewers conducting Spanish interviews with poor Spanish fluency. On the whole, each of the different pretesting methods uncovered different types of issues and/or reinforced findings from other methods. They all helped 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.

Ideal Sequence for Multiple Pretesting Methods in the Development of a Bilingual Instrument. We recommend that prior to any field testing, translations be thoroughly reviewed using the committee approach (See U.S. Census Bureau, 2004). The next step should be concurrent iterative rounds of cognitive and usability testing of both language versions (See Goerman & Caspar, 2007, for more information on this type of cognitive research). 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 and usability testing. This recommended timeline for pretesting would allow for different types of improvements to be made to the questionnaires at each stage. These new wordings 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 five 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 will "work" in the field. We examined the survey from the interviewer perspective via the usability and observational studies 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 is unique in that it was done in two languages, which enabled us to examine equivalency of meaning and interpretation across the source and translated versions of an instrument in each of these steps.

Acknowledgements

We would like to thank Theresa DeMaio, Yuling Pan, Stephanie Sheffield, and Nathan Jurgenson for reviewing earlier drafts of this paper and providing insightful comments.

References Cited

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., 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., 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. and Landreth, A. (2006). Analyzing Interviewer/Respondent Interactions while Using a Mobile Computer-Assisted Personal Interview Device. Field Methods, Vol. 18, No. 3, 335-351.

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.

Dumas, J. S., and Redish, J. C (1999) *A Practical Guide to Usability Testing (Revised Edition).* Exeter, UK: Intellect.

Fowler, F. (1992). How unclear terms affect survey data. *Public Opinion Quarterly, 56,* 218–31.

Fox, R. (1991). *Encounter with Anthropology*. Edison, NJ: Transaction Publishers.

Goerman, P. (2005). Making ends meet: The complex household as a temporary survival strategy among new Latino immigrants to Virginia. In Shwede, L., Blumberg, R. L., and Chan, A. Y. (Eds). *Complex Ethnic Households in America.* 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. (2007). "A New Methodology for the Cognitive Testing of Translated Materials: Testing the Source Version as a Basis for Comparison." Paper presented at the American Association for Public Opinion Research conference, May 17 - 20, 2007, Anaheim, California. *2007 JSM Proceedings, Statistical Computing Section [CD-ROM]*, (pp 3949–3956). Alexandria, VA: American Statistical Association.

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

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., Lepkowski, J.M. (eds.) Hyattsville, MD: National Center for Health Statistics, 2004.

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 A. D. Landreth. (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.

Love, K., and Byrne, R. (2005). Investigating Measurement Errors Due to Interviewer Misclassification and Mode. DSSD 2010 Census Coverage Measurement Memorandum Series #2006-D7-03.Unpublished internal report for the U.S. Census Bureau.

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.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New strategies for pretesting survey question. *Journal of Official Statistics, 7,* 349–394.

Olmsted, E. and Hourcade, J. P. (2005). NRFU Round II Quick Report. Submitted to DMD August 25, 2005.

Olmsted, E., Hourcade, J. P., and Abdalla, A. (2005). NRFU 2006. Round I Quick Report. Submitted to DMD August 8, 2005.

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, 2004.

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 Conty, Texas Spring and Summer 2006. Report submitted by Development Associates to the Census Bureau, December 4, 2006.

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

U. S. Census Bureau (2003). Census Bureau standard: Pretesting questionnaires and related materials for surveys and censuses. U.S. Department of Commerce, Washington D.C.: Author.

Appendix A. Draft of Revised Information Sheet for 2010 Census NRFU.

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 for no other purpose. As allowed by law, your census data become 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/privacy/> for additional information.

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) (6-17-2008)

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 these people: (They will be counted at the other place)

- College students who live away from this address most of the year
- Armed forces personnel who live away
- People who, on April 1, 2010, were in a:
  - Nursing home, mental hospital, etc.
  - Jail, prison, detention center, etc.

INCLUDE these people:

- Babies and children living here, including foster children
- Roommates
- Boarders
- People staying here on April 1, 2010 who have no other 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, of 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) (6-17-2008)