Explaining Differences in Inter-coder Reliability between English and Spanish Language Behavior Coding Research[1]

Patricia L. Goerman, Jennifer H. Childs and Matthew Clifton
Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100

Abstract

Previous behavior coding studies have shown that there is often variation in inter-coder reliability across different language versions of a survey instrument (Edwards et al., 2004; Hunter and Landreth, 2006). This paper examines the issue of inter-coder reliability in multilingual behavior coding studies with a focus on bilingual (Spanish/English) behavior coding research conducted by the U.S. Census Bureau. During the past decade the Census Bureau has worked to develop a questionnaire for use in follow-up interviews for non-respondents in the 2010 Census.[2] As a part of the development process, early versions of the survey instrument were field tested and behavior coded in both 2004 and 2006. For both studies, a sample of Spanish and English language interviews were tape recorded. Behavior coding examined both interviewer and respondent behaviors in order to assess and improve question wording. Within each study, the same bilingual coders were used to code both Spanish and English cases (though the coders differed across studies). Inter-coder reliability was calculated based upon all of the coders having coded the same set of three or four cases in English and three or four cases in Spanish each year. In 2004, we found that inter-coder reliability was consistently lower in Spanish than in English. In 2006, inter-coder reliability was relatively low overall with some variation across languages, with coders doing better in some respects in English and in other respects in Spanish. Through an examination of notes taken by the coders in both studies and in methods used in each study, this paper aims to explain differences in reliability between the Spanish and English cases and to make recommendations as to how to improve the methods used when coding interviews in a multilingual behavior coding study.

Key Words: Bilingual behavior coding, inter-coder reliability, pretesting survey translations, Spanish-language survey research

1. Introduction

With increases in globalization and migration in the United States and around the world, survey research organizations are increasingly working to translate questionnaires and other survey materials into multiple languages. Along with the increased need for survey translation, researchers are recognizing the need to pretest survey translations in order to ensure that parallel data are being collected across language groups. A number of pretesting methods have been adapted and used to aid in this process. There has been a great deal of research on the use of expert review in the form of committee approaches to translation (U.S. Census Bureau, 2004; Harkness, et al., 2004; Harkness, 2002) and on the use of cognitive testing to test survey translations (Goerman and Caspar, 2007; Pan, 2004). This paper focuses on the application of another pretesting method to the evaluation of survey translations: behavior coding. Very little research has been done regarding how to best apply this method to the evaluation of multiple language versions of a survey instrument. This paper provides some lessons learned from bilingual (Spanish/English) behavior coding projects at the U.S. Census Bureau.

Since 2004, pretesting of survey translation has been an area of focus for the Census Bureau (Gerber and Pan, 2004). The Census Bureau (2004) released translation guidelines that recommend pretesting all survey translations for "semantic, conceptual, and normative equivalence." Additionally, the Census Bureau has a Pretesting Standard which requires that all survey questions be pretested and shown to "work" prior to being fielded (U.S. Census Bureau, 2003). These standards and guidelines recommend pretesting questions in the languages in which they will be administered. In addition to focus group research and cognitive interviewing (Goerman 2006; Pan 2004), behavior coding has recently been used as another method of pretesting survey translations at the Census Bureau. However, research on best practices for the implementation of bilingual behavior coding is quite limited.

2. Review of the Literature on Behavior Coding

Behavior coding was first developed in the 1960s and is an interaction-based approach in which interviewer and respondent behaviors are observed during the administration of survey questions (Cannell, et al. 1968). Researchers then document the way in which a survey was actually carried out compared to the ideal administration as envisioned by survey designers (Fowler and Cannell, 1996). Different types of interactions are assigned specific codes which are applied and analyzed to provide a quantitative means of identifying problematic survey questions (Sykes and Morton-Williams, 1987). The results of

[1] *Disclaimer:* This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] Originally, the Census Bureau was going to use a Computer Assisted Personal Interview (CAPI) instrument to collect non-response follow-up data through personal visits. This paper describes the testing of that instrument. However, since then, a decision was made to use a paper questionnaire for non-response follow-up data collection in 2010.

the analysis of interviewer and respondent behavior obtained through behavior coding are used to provide input for the rewording or redesign of survey questions. Many researchers laud behavior coding as being extremely useful in that it provides for a quantitative analysis of verbal behavior, something which does not lend itself easily to mathematical summarization (Burgess and Paton, 1993).

2.1 The Use of Behavior Coding in Pretesting Survey Translations

Behavior coding is particularly useful for pretesting a survey translation, in that it enables researchers to identify problematic translations through the analysis of respondent behaviors, such as a request for clarification or a declaration of confusion. Just as researchers often recommend the cognitive testing of the original source language version of a survey instrument along with the translated version (Goerman and Caspar, 2007), we argue that it is advisable to conduct concurrent behavior coding of source and translated instruments in order to identify places where interviewers or respondents have different levels of difficulty in administering or answering questions. This can indicate a place where a translation is not conveying the same meaning as the source question and the results can guide researchers in making revisions to question wording.

2.2 Inter-coder Reliability in Behavior Coding Research

Inter-coder reliability is a measure of the extent to which two or more coders agree when applying a coding scheme to a given set of information. There are many different statistical measures often used to examine inter-coder reliability and there is no consensus on which is the best measure (Cho, 2008). However, to reduce the likelihood of coder bias negatively affecting behavior coding results, Fowler and Cannell (1996) recommend use of the kappa statistic. This measure is commonly used in behavior coding research. The kappa statistic is a measurement of agreement between two or more sources that reduces the likelihood of agreement through mere chance (Fleiss, 1981). By measuring reliability, researchers can have greater confidence in the analyses of interviews by coders in behavior coding research.

3. Background on Bilingual Behavior Coding Projects at the Census Bureau

The Census Bureau conducted its first bilingual behavior coding research as a part of the development of a non-response follow-up instrument for use in the 2010 Census (Childs, et al., 2007; Hunter and Landreth, 2006). Field tests were conducted in 2004 and 2006 using draft versions of the instrument in both English and Spanish. The Census Bureau conducted behavior coding in order to evaluate draft question wording in both languages, and revisions were made based on the results of each study. During both field tests, observers accompanied interviewers and asked respondents' permission to tape record their interviews. Those interviews were later coded by bilingual coders and the resulting codes were analyzed by researchers. Results and lessons learned from the 2004 study informed the methods used in the 2006 study. This paper discusses lessons learned from both studies and makes recommendations for future research and modification to the methods.

The survey questions tested in the two studies were basic demographic census questions. There were several household–level questions which asked for the number of people living in the household, whether the home was their usual place of residence, and whether the home was rented or owned. There were also a number of person–level questions that asked for the name of each household resident, the relationship between them, sex, age, date of birth, Hispanic origin, and race.

There are some limitations that should be considered when interpreting the results of these studies. First of all, due to our method of gathering tape-recorded interviews, we were unable to draw a random sample of households. Secondly, both interviewers and respondents were aware that they were being tape recorded, which may have caused them to behave in an unnatural manner. Despite this fact, interviewers and respondents did exhibit surprisingly high rates of undesirable behavior.[3]

4. Methodology of Census Bureau Behavior Coding Projects

In 2004 a total of 220 taped interviews were gathered by Census Bureau staff, exceeding the target of 200 interviews; the breakdown of cases by language is presented in Table 1. Thirty-six cases were not usable for reasons such as problems with the equipment, inadvertently taped proxy interviews or the mistaken omission of the respondent giving consent to be recorded on the tape itself.

Unfortunately, there were problems in 2006 when the Census Bureau contracted the tape recording to an outside company. The target was to tape 120 interviews, 60 in each language, however there were a total of only 72 interviews that could be analyzed in 2006; 54 of these were in English and only 18 were in Spanish. Twenty-nine additional cases were tape recorded but had to be discarded, mostly due to an error that resulted in not having recorded respondent consent on the tapes. We believe that the small number of cases in 2006 had adverse effects on our results, which we discuss further below.

[3] An ideal way to deal with these issues would be to have recording equipment incorporated into a CAPI instrument that could record cases at random without interviewer awareness. Unfortunately, we did not have access to this type of technology at the time of these studies.

Table 1: Number of Tape Recorded Behavior Coding Cases in Census Studies

2004	2006
220 Interviews analyzed	72 Interviews analyzed
119 English	54 English
72 Spanish	18 Spanish
29 Part English/ part Spanish	
36 Not usable	29 Not usable
46 Interviewers recorded	22 Interviewers recorded

4.1 Coder Selection, Characteristics, and Training

In the 2004 study, five bilingual behavior coders coded both the English and Spanish language cases. The coders varied by a number of characteristics (See Table 2). The coders in both studies were Census Bureau telephone interviewers who had been recommended by their supervisors as fully bilingual and as skilled interviewers. They ranged from having one year to two and a half years of experience as telephone interviewers. The 2006 study also began with a total of five coders, but two of the coders were unable to participate in the project after the training session had been completed.

Table 2: Coder Selection and Characteristics

2004	2006
5 Bilingual coders	3 Bilingual coders
3 Native Spanish speakers	2 Native Spanish speakers
2 Native English speakers	1 Native English speaker
3 Born in U.S., 2 born in Mexico	2 Born in U.S., 1 born in Mexico
1-2.5 years of interviewing experience	1.5-8 years of interviewing experience

Coder training was conducted in a very similar manner across the two studies. For each year, the training consisted of a three-day session, conducted primarily in English because two of the lead researchers were monolingual English speakers. There was one bilingual Spanish-speaking lead researcher each year who led the Spanish language parts of the training. Training consisted of theoretical background on behavior coding methods and on the particular coding scheme that was used in each project. It also included extensive coding practice and feedback. In 2004, the lead researchers had observed interviews in the field and they created dramatized, tape-recorded interactions to illustrate the various codes. Coders listened to the tapes and practiced assigning codes to the interactions. They were then provided with feedback on their coding. Because this method was time-consuming and may not have provided realistic examples of all types of coding situations, a change was made to the 2006 training methodology. In 2006, the lead researchers listened to interview tapes and identified sample interactions that illustrated various codes. The coders listened to these real interactions and practiced assigning them codes. Next the coders practiced coding an entire interview in English and an interview in Spanish, followed by feedback and discussion.

Another difference between the methodologies of the two studies was that in 2004 the coders had to record codes on paper coding sheets. In 2006, we created a computer database for coding. The database increased the ease of the training session in that practice codes could be quickly analyzed and compared across cases to see where coders were having difficulty.

4.2 Behavior Codes

The behavior codes used in these two projects were adapted from Oksenberg, Cannell, and Kalton's (1991) research. Some changes were made between the two studies based on codes that had caused confusion in the 2004 research. The codes covered three basic types of situations: 1) Question-asking behavior by the interviewer on the first level of exchange[4]; 2) Response behavior by respondents on the first level of exchange; and 3) Final outcome, defined as the ultimate response, or lack thereof, that was agreed upon by the interviewer and respondent at the end of their interaction.

An additional methodological decision made in both of these projects was to have coders take extensive notes about any problematic interaction that occurred. This enabled the researchers to analyze how and why coders had erred in coding.

4.2.1 Interviewer Behavior Codes

In terms of interviewer behavior, the codes used were the same in 2004 and 2006. Coders listened to the interviewers' administration of each question and coded them with six possible options. Questions were to be coded as an "Exact reading/slight change" when an interviewer read a question exactly as worded or made only a slight modification that would

[4] For the purposes of these studies, the first level exchange was defined as encompassing the first thing that the interviewer said without interruption followed by the respondent's first "turn" speaking in the interaction.

not affect question meaning. A code called "Major change" was to be used when interviewers made a major change to the question wording that could impact respondents' understanding of the question. The third code was a "Correct verification" which occurred when an interviewer verified something that a respondent had already said during the course of the interview. Census interviewers are allowed to verify this type of information rather than asking the interview question exactly as worded. For example, an interviewer might say "You said that Juan is your husband, right?" The fourth code was "Incorrect verification" which was to be used when an interviewer verified information that was not based on something that the respondent had said during the recorded interview. One limitation to this code is that it is possible that a respondent had provided information to the interviewer prior to initiating the tape recording. Coders were instructed to list an interaction as "Inaudible/Other" if they had trouble hearing or understanding it on the tape. In some cases this may have been due to gestures or other communication that would not be picked up on an audio tape. Finally there was a code for questions that were "Incorrectly skipped" by interviewers. The only two codes considered to be "good" interviewer behavior were "Exact reading/slight change" and "Correct verification."

4.2.2 Respondent Behavior and Final Outcome Codes

When coding respondent behavior, we started with Oksenberg, Cannell and Kalton's basic codes in the 2004 study but made some modifications prior to the 2006 study. Coders were instructed to apply the "Adequate answer" code whenever a respondent gave a codeable response that fit the response categories as his/her first utterance after hearing the interviewer read the question. The "Inadequate answer" code was to be assigned when a respondent said something that could not be easily coded in the instrument. The "Qualified answer" code was to be used when a respondent provided a codeable answer but then added additional information that might change the initial response if factored in. For example, if the interviewer asked "How many people live or stay here?" and the respondent said "Three, if you don't count my college student," this would be a qualified answer. If in response to the same question, the respondent said "I'm not quite sure if you want me to count my college student. I guess I'll say three," this should be coded as an "Uncertain answer." Additionally, a respondent could ask for more information prior to answering the question and this would be coded as a "Request for clarification." Finally, the "Request for a re-read" code was to be used when a respondent asked the interviewer to repeat the question. There were also standard "Don't know" and "Refusal" codes.

In analyzing the 2004 data, the researchers noted that there was confusion on the part of the coders regarding the difference between the "Uncertain answer" and the "Qualified answer" codes. Additionally, researchers determined that it was not of particular use to distinguish between "Requests for clarification" and "Requests for a re-reading" of a question. In analyzing coder notes in cases where inter-coder reliability was low, the researchers realized that many of the interactions could legitimately have been assigned more than one code, so each of these sets of two codes were collapsed into one for the 2006 study. As a result, there were six respondent codes instead of eight codes in the 2006 study. The researchers hoped that this would improve inter-coder reliability.

The final outcome was coded in each study as well. Final outcome codes were the same as Respondent behavior codes except that the "Request for clarification" and "Request for re-read" codes were omitted, since an interaction would not normally end with a request for additional information.

5. Inter-coder Reliability

When conducting a behavior coding study it is important to incorporate a measure to evaluate the accuracy of the coding of interviewer and respondent behavior. Calculating inter-coder reliability helps determine whether individual coders have applied the same codes to each type of interaction. In the event that the same codes have not been applied, it can be difficult for researchers to determine which code is correct. Coder notes on problematic interactions between interviewers and respondents allow researchers to further investigate unreliably coded interactions.

Confidence in the reliability of the coders in the two Census studies was of utmost importance in interpreting whether the survey questions were working well or whether they were in need of revision. In order to examine this issue, we had each of the coders code some of the same cases so that we could assess their level of agreement in assigning codes. In 2004, each of the five coders coded the same four English and four Spanish cases. In 2006, each of the three coders coded the same three English and three Spanish cases.

In order to measure inter-coder reliability we used the Kappa statistic. Following other behavior coding studies, we considered Kappa scores of .75 and greater to indicate excellent agreement among coders. Scores between .40 and .75 were considered to have a good to fair level of agreement. Finally, reliability scores of .40 and lower were considered poor agreement among coders.

5.1 Results by Language

In both of these projects the same coders worked on Spanish and English cases, and interestingly, we found that inter-coder reliability varied across languages. The next sections provide results and discussion of inter-coder reliability in each study.

5.1.1 Inter-coder Reliability in the 2004 study

In 2004, each of the five bilingual coders coded the same four English and four Spanish cases. Interestingly, the same coders agreed more often when coding the English cases than the Spanish cases (See Table 3). During an initial review of the dataset by the researchers, it was evident that coders had some difficulty applying the codes in both languages, and that they often coded the same interaction in different ways. For example, the coders sometimes mistakenly coded what should have been considered an interviewer "verification," such as, "You're Hispanic, right?," as a "major change." Using the coder notes, the researchers recoded cases where it was clear that the coders had made mistakes in applying the coding scheme. This step was undertaken prior to completing our analysis of the results and prior to making recommendations on question wording.

Table 3: Inter-coder Reliability by Language, 2004

	Interviewer Behavior	Respondent Behavior	Final Outcome
English	.63 (Good)	.48 (Fair)	.70 (Good)
Spanish	.50 (Fair)	.34 (Poor)	.31 (Poor)

The fact that the Spanish cases resulted in poorer reliability was surprising given that the same coders were working across languages. The researchers hypothesized that there may have been an interaction between translation quality, the language skill of interviewers, the language skill of coders, the comparative complexity of the Spanish interviews versus the English interviews, and the sometimes legitimate assignment of different codes to the same interaction. However, without further analysis of the taped interviews, it was impossible to know the exact cause of differences in reliability across languages. Because of these differences in reliability, the researchers decided to collapse some of the codes and use real-life Spanish examples in the training to see if the results could be improved in the 2006 study.

5.1.2 Inter-coder Reliability in the 2006 study

In 2006, the three bilingual coders coded the same six interviews, three in English and three in Spanish. Interestingly, the 2006 Spanish language scores were similar to the 2004 Spanish scores, but the 2006 English scores went down compared to the 2004 English scores (See Table 4). The researchers attributed most of the low English scores to the fact that one coder had a particularly steep learning curve. In addition, in 2006 there were only three coders as opposed to the five coders in 2004. All of the coders had started coding English cases prior to Spanish cases, and one of them had a particularly hard time at first. Her coding improved as the project progressed. An additional explanation for this result may be that two of the three coders were native Spanish speakers and may not have been as proficient in English as the researchers would have liked.
As a result of the low reliability and evidence of coding errors, the researchers again had to go back and recode much of the data before analyzing and presenting the results on each survey question. The detailed coder notes made this feasible.

Table 4: Inter-coder Reliability by Language, 2006

	Interviewer Behavior	Respondent Behavior	Final Outcome
English	.34 (Poor)	.40 (Fair)	.52 (Fair)
Spanish	.49 (Fair)	.39 (Poor)	.48 (Fair)

5.2 Qualitative Analysis: Coder Notes

The use of coder notes is somewhat uncommon in behavior coding studies but is routinely done at the Census Bureau. For the bilingual studies, coders had been asked to take detailed notes on the interactions any time an interviewer or a respondent displayed less than ideal behavior. Having these notes was useful in that researchers could look in detail at the coders' impression of an interaction to see why they may have coded something differently from each other. We could also determine when coders had coded things contrary to their training. Finally, we were able to look in-depth at the differences between the Spanish and English cases. Through analysis of the coder notes we found a number of reasons that reliability was often lower in Spanish language cases. The next section provides two examples of interviewer-respondent interactions that illustrate the complexity of the coding task.

5.2.1 Spanish Language Coding Difficulty

One of the questions tested in the 2006 study was the "Hispanic origin" question, which reads: "Are you/ is NAME of Hispanic, Latino, or Spanish origin?" The following was an exchange in one of the Spanish language Kappa interviews (translated into English):

> Interviewer: Is Ana of Hispanic, Latino, or Spanish origin?
> Respondent: What should I put there? Because she's Mexican.
> Interviewer: She's not Hispanic, Latino, or Spanish?

Respondent: Spanish? No.
Interviewer: Ok, so “no.”

This case was coded by all three coders and the bilingual lead researcher. In terms of interviewer behavior, all four coders assigned the “Exact reading/slight change” code to this interaction. Respondent and final outcome behaviors were not as straightforward. For respondent behavior, three of the coders chose “Inadequate answer” and one chose “Request for clarification.” The researchers noted that both of these codes do apply in this case, although coders should have chosen the more specific “clarification” code based on their training. Finally, for final outcome behavior, three coders classified the final “no” response as an “Inadequate answer” and one classified it as an “Adequate answer.” The correct answer was actually “Adequate answer,” since this was a codeable response that fit the response categories, and it appears to be the answer that both interviewer and respondent agreed upon. This example illustrates a potential problem with our coding scheme, as coders were understandably reluctant to classify this as an “adequate” answer when it so clearly seemed to be an incorrect answer.

5.2.2 English Language Coding Difficulty

While not as common, problems occurred in the English-language coding as well. When coding a response to the same Hispanic origin question above on a Kappa case, an interviewer-respondent interaction went as follows:

Interviewer: Are you of Hickspanic [sic], Latino, or Spanish oregon [sic]?
Respondent: I’m black.

This exchange was followed by laughter on the part of both the interviewer and the respondent. This case was coded by all three coders and the three lead researchers. In terms of interviewer behavior, five people coded this as “Exact reading/slight change” and one person coded it as “Major change.” The interviewer made some pronunciation mistakes that may or may not have changed the meaning of the question. This was not something that we had thought to address in the coder training. On the whole, pronunciation difficulties were much more common in the Spanish language interviews and there was often evidence that this may have contributed to difficulty in coding. In terms of respondent and final outcome behavior, all six coders designated this exchange as yielding an “Inadequate answer.” Because “I’m black” is not one of the response options that can be coded for this question, this response was very straightforward and easy for all coders to classify.

6. Explaining Differences in Reliability across Languages

Based on our analysis of coder notes, we identified a number of possible reasons that coding Spanish cases seemed to be more difficult for coders. Table 5 shows the percent of “good” behavior for interviewers and respondents from the 2006 data.[5] Interviewer behavior was classified as “good” when it was assigned either an “Exact reading/Slight change” rating or a “Correct verification” rating. Respondent and final outcome behaviors were considered “good” when they were assigned an “Adequate answer” code.

Table 5: Percentage of “Good” Behavior Based on Recoded 2006 Data

	Interviewer Behavior	Respondent Behavior	Final Outcome
English	45%	82%	89%
Spanish	31%	69%	79%

On the whole, interviewers asked questions as they should have more often in English than in Spanish. Similarly, respondents gave codeable answers on the first exchange and reached a codeable final outcome more often in English. The effect of language was statistically significant (Childs et al., 2007). Because there are only one or two codes required to describe good behavior and many more codes that can be assigned to less desirable behavior, and because the English language cases more often resulted in good behaviors, the English language interviews should have been easier to code.

A second area that we identified which may have contributed to lower inter-coder reliability in Spanish was that Spanish language interviewers were observed to be variably fluent in the Spanish language. Just as our coders were not tested for proficiency in Spanish and English, the field interviewers employed for the field test had not been screened for fluency prior to being allowed to conduct interviews in Spanish. When reviewing the tapes, we noticed that interviewers sometimes had difficulty reading the Spanish questions aloud as worded. Some of them had difficulty pronouncing certain words and in some cases we even observed respondents correcting interviewer pronunciation. As one of the examples above illustrates, it can be difficult to code an interaction in which poor pronunciation or fluency might cause a question meaning to change. We observed this to be a factor more often in the Spanish language cases than in the English cases.

A third issue is that the specific cases chosen to measure reliability in Spanish may have been more difficult to code than the English ones. Some interviews contain more discussion, involve larger households and are with less-educated respondents

[5] The designation of “good” behavior is based on re-coded data which was revised by the researchers based on the coder notes.

than others. Any of these factors can cause an interview to be lengthy or an interaction to be less straightforward. It can be a challenge for researchers to evaluate the difficulty level of coding particular cases and to provide cases of a similar difficulty level across languages both for training and for the purposes of reliability measurement.

A fourth issue that may have affected differences in inter-coder reliability was that certain questions may be more or less difficult for respondents of a certain language group to understand or answer. The Hispanic origin question described above proved particularly difficult for Spanish-speaking respondents to answer. We observed confusion over whether the question was asking people if they were "Spanish" meaning "from Spain," whether it was asking them to choose one of the three "options" (Hispanic, Latino or Spanish), and whether it was asking for specific national origin as opposed to being a "yes or no" question. Any complications for a particular group of respondents will make coding those cases more difficult.

An additional issue that we encountered was that the Spanish-language instrument had not been thoroughly reviewed or cognitively pretested prior to its fielding. Because of this, the instrument included terms that did not work well in Spanish and typos, which made for more difficult interviewer administration. Both of these things may have increased respondent confusion and interviewer burden. These issues also most likely created a more difficult coding task for the behavior coders.

Finally, through our analysis of the tapes and coder notes, we realized that in some cases different codes could legitimately be assigned to the same interaction. Our codes were not completely mutually exclusive. The above example about whether pronunciation errors should be coded as a "Major change" or an "Exact reading/Slight change" illustrates this type of situation.

7. Conclusion: Methods for Improving Inter-Coder Reliability in Bilingual Behavior Coding

This research has led us to a number of next steps that we would recommend for improving methods in future bilingual behavior coding studies. First of all, we recommend a simplification and reworking of the behavior coding scheme to be sure that codes are as mutually exclusive as possible. One problem that we noticed in particular was that the "Adequate answer" as a final outcome code was misleading for many coders. It was sometimes clear to the coders that while a final, codeable answer was obtained, it did not seem to be the correct answer. The Mexican respondent who was recorded as not being Hispanic is a perfect example. We recommend either rewording the "Adequate Answer" category to read "Codeable Response" and/or adding another category to keep track of instances when an incorrect answer seems to have been recorded in the instrument. In many cases this would be a subjective measure but it still might provide interesting and useful data.

Other methods that we recommend for improving inter-coder reliability in bilingual behavior coding projects include a thorough review of translated instruments, including improvements through cognitive testing prior to fielding an instrument for behavior coding. We also recommend that both field interviewers and coders be certified as to their proficiency in the languages in which they will be working prior to embarking on a project of this nature.

Coder selection and training should also include the hiring of more coders than are ultimately desired as a minimum number since personnel changes can occur through the course of a project. Training itself should be in-depth and should include as much practice work with feedback as possible. Coders' work could conceivably be monitored and evaluated during the coding process as well so that errors in application of codes could be caught early and corrected.

Finally, we recommend that researchers work to ensure that cases selected for reliability coding have similar characteristics, (e.g., length of interview), so that there will not be a large difference in ease of coding across languages. This may be difficult to achieve as interviews can be more difficult for both interviewers and respondents in different languages.

8. Areas for Future Research

These studies have brought up a number of areas for which we recommend further research. We recommend the conduct of methodological research into whether coding the same data with different methods would provide improved reliability. For example, the same set of interviews could be coded by different coders after undergoing variations in training, using different coding schemes and/or using coders with different characteristics to see how reliability can best be achieved across languages. Along these same lines, an interesting area to investigate would be whether it is better to have the same coders code in both languages or whether one can achieve better results by having native speakers code only in their dominant language, thus using separate coders for each language.

For these two Census projects, we did not link up the responses that were actually recorded in the electronic survey instrument to see if they matched what appeared to happen in the verbal discussions between interviewers and respondents on the audiotapes. This was due to time constraints in the initial projects. We recommend that future projects incorporate a linkage between verbal behavior and what is recorded by interviewers, as interviewers may sometimes compensate for respondents' poor understanding of questions by taking it upon themselves to answer differently.

Acknowledgements

We would like to acknowledge the researchers who worked on the two original Census Bureau bilingual behavior coding projects: Ashley Landreth, Carlos Arce, Dawn Norris and Aref Dajani. We would also like to thank Theresa DeMaio, Yuling Pan and Tommy Wright for reviewing this paper and Stephanie Sheffield for editorial assistance.

References

Burgess, M. J. and Paton, D. (1996). "Coding of Respondent Behaviour by Interviewers to Test Questionnaire Wording." Paper submitted to *1993 ASA Proceedings, Section on Survey Research Methods, Vol. 1 ,* Alexandria, VA: American Statistical Association: pp-pp. 392-397.

Cannell, C., Fowler, F., and Marquis, K. (1968). "The Influence of Interviewer and Respondent Psychological and Behavior Variables on the Reporting in Household Interviews." *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research* (26), pp. 1-65..

Childs, J.H., Landreth, A., Goerman, P., Norris, D., and Dajani, A. (2007). "Behavior Coding Analysis Report: Evaluating the English and the Spanish Version of the Non-Response Follow-Up (NRFU)." *Statistical Research Division Report Series (Survey Methodology # 2007-16).* U.S. Census Bureau. On-line, available: http://www.census.gov/srd/papers/pdf/ssm2007-16.pdf.

Cho, Y.I. (2008). "Intercoder Reliability." In Lavrakas, P.J. Encyclopedia of Survey Research Methods, Volume 1. (pp. 344-345). Thousand Oaks, CA: Sage Publications, Inc.

Edwards, W. S., Fry, S., Zahnd, E., Lordi, N., Willis, G., Grant, D. (2004). "Behavior Coding across Multiple Languages: The 2003 California Health Interview Study as a Case Study." Paper presented at the American Association for Public Opinion Research conference, May 13-16, 2004, Phoenix, AZ, and submitted to *2004 ASA Proceedings [CD-ROM]*, Alexandria, VA: American Statistical Association: pp. 4766 – 4774.

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

Fowler, F. J. (1992). "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly*. 56(2), pp. 218-231.

Fowler, F., and Cannell, C. (1996). "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." Schwartz, N., and Sudman, S. (Eds.) Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. (pp. 15–36). San Francisco: Jossey-Bass.

Gerber, E. and Pan, Y. (2004). "Developing Cognitive Pretesting for Survey Translations". Paper presented at the Second International Workshop on Comparative Survey Design and Implementation (CSDI), April 1-3, 2004. Paris, France.

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. On-line, available: http://www.census.gov/srd/papers/pdf/rsm2006-03.pdf.

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 and submitted to *2007 JSM Proceedings, Statistical Computing Section [CD-ROM]*, Alexandria, VA: American Statistical Association: pp-pp. 3949 – 3956.

Harkness, J., Pennell, B., and Schoua-Glusberg, A. (2004). "Survey Questionnaire Translation and Assessment." In Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., and Singer, E. (Eds.), Methods for Testing and Evaluating Survey Questionnaires (pp. 453-474). Hoboken, NJ: John Wiley and Sons, Inc.

Harkness, J. (2002). "Questionnaire Translation." In Van De Vijver, F.J.R., Mohler, P.Ph. and Harkness, J.A. (Eds.), Cross-Cultural Survey Methods (pp. 35-56). Hoboken, NJ: John Wiley and Sons, Inc.

Hunter, J. and Landreth, A. (2006). "Behavior Coding Analysis Report: Evaluating Bilingual Versions of the Non-Response Follow-Up (NRFU) for the 2004 Census Test." *Statistical Research Division Report Series (Survey Methodology #2006-07).* U.S. Census Bureau. On-line, available: http://www.census.gov/srd/papers/pdf/ssm2006-07.pdf.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). "New Strategies of Pretesting Survey Questions." *Journal of Official Statistics*, 7(3) pp. 349-366.

Pan, Y. (2004). "Cognitive Interviews in Languages other than English: Methodological and Research Issues." Paper presented at the American Association for Public Opinion Research conference, May 13-16, 2004, Phoenix, AZ, and submitted to *2004 ASA Proceedings [CD-ROM]*, Alexandria, VA: American Statistical Association: pp. 4859 - 4865.

Sykes, W. and Morton-Williams, J. (1987). "Evaluating Survey Questions." *Journal of Official Statistics*. 3(2) pp. 191-207.

U.S. Census Bureau (2003). *Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses.* Methodology and Standards Directorate, U.S. Census Bureau, U.S. Department of Commerce, Washington D.C. On-line, available: www.census.gov/srd/pretest-standards.pdf.

U.S. Census Bureau (2004). *Census Bureau Guideline: Language Translation of Data Collection Instruments and Supporting Materials.* U.S. Census Bureau, U.S. Department of Commerce, Washington, D.C. On-line, available: http://cww.census.gov/msdir/docs/G08-0_Language_Translation.pdf.