Exploring the Reliability of Behavior Coding Data

Nathan Jurgenson[1] and Jennifer Hunter Childs[1]
[1]Center for Survey Measurement, U.S. Census Bureau,
4600 Silver Hill Rd. Washington, DC 20233

Abstract[1]
Behavior coding examines interviewer and respondent interactions, often with the intent of evaluating the quality of survey questions. In practice, several coders listen to these interactions and code prescribed behaviors. Often, the coders are asked to code the same cases to be used as a measure of reliability of the coding. If the reliability is poor, then the results of the coding may be less meaningful. The kappa statistic, a conservative measure of reliability that corrects for chance agreement, is often used. This paper explores different ways of measuring reliability using the kappa statistic and what can be learned by examining reliability of the codes in different ways. For example, we will examine reliability scores by speaker (e.g., interviewer, respondent), by question (e.g., age, race), by code (e.g., exact question reading, request for clarification) and by type of question (e.g., select one, multiple choice). Examining reliability in these ways will allow us to recommend a strategy for reporting reliability in behavior coding studies as well as learn more about the behavior coding method itself and characteristics of questions that can cause problems.

Key words: Reliability, Behavior Coding, Questionnaire Pretesting and Evaluation, Kappa

1. Introduction

This paper presents results of an exploratory study on the reliability of behavior coding data of survey interviews. Behavior coding is often seen as an important survey evaluation method (Willis 2005). However, the usefulness of this method depends on the reliability of the coders. For this study, we look at the coding of English and Spanish language interviewer and respondent behavior using the 2010 Census Nonresponse Followup (NRFU) interviews. The kappa statistic is often used to measure reliability. It is a conservative measure that corrects for agreement by chance. This paper explores different ways of measuring reliability using the kappa statistic and what can be learned by examining reliability of the codes in different ways. We look at the relationship between coder reliability and question length, the accuracy of the interviewer's reading or respondent's response, and the type of response the question requires. Importantly, we also discuss behavior coder reliability with respect to English and Spanish language interviews. Examining reliability in these ways will allow us to recommend a strategy for reporting reliability in behavior coding studies as well as learn more about the behavior coding method itself and characteristics of questions that can cause problems for interviewers, respondents and behavior coders.

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

1.1 What is Behavior Coding?

The behavior coding method is used in survey research to analyze the interactions between interviewers and respondents during the administration of survey questions (Cannell, Fowler, and Marquis, 1968). As the name suggests, the method involves the systematic application of codes to behaviors (in this case, verbal behaviors) that interviewers and respondents display during the question/answer process and is often used to identify problematic questions (Oksenberg, Cannell, and Kalton, 1991; Sykes and Morton-Williams, 1987).

Behavior coding is a useful method for gathering information about the quality of the survey questionnaire and the data it collects. If questions and response options are worded and structured in ways that respondents can easily understand and answer, then confidence grows regarding the ability of the survey questionnaire to meet its intended measurement objectives. In an ideal interaction between an interviewer and a respondent, the interviewer asks the question exactly as worded and the respondent immediately provides an answer that is easily classified into one of the existing response categories. When the interaction deviates from this ideal, we begin to suspect there may be problems with the question and/or response options that may be causing comprehension or response difficulties. These difficulties could lead to measurement error. The application and analysis of behavior codes for these types of interactions allow researchers to pinpoint where such issues are occurring in the survey questionnaire (Fowler and Cannell, 1996).

Interviewers are trained to administer the survey in a specific way, and behavior coding allows researchers to observe how the interviewer(s) actually accomplish their task by systematically applying various codes to their behavior. Further, behavior codes can be applied to the respondent side of the interaction as well. Importantly, the behavior coder is usually not present during the interview.[2] Instead, the interviews are often audio-recorded so the presence of the coder does not alter the interview process. It should be noted, however, that the presence of a recording device may influence the question asking and responding processes.

The codes themselves serve as a description of the interaction process and coders are trained to identify which codes should be placed on which behaviors. As such, behavior coding helps transform observable interview behaviors into quantitative tallies. The sorts of codes depend on the specific project, but usually describe whether the interviewers read the questions, if they read them correctly, and if they made any major changes to question wording. Researchers may also employ codes that document if the respondent easily answered the question or if the respondent had difficulty answering, often demonstrated by requests for clarification, answering outside of the response set or refusing to answer.

The advantages of the systematic coding of interview behaviors are clear. Researchers can learn more about how interviewers are performing their task and how respondents answer to best determine if the questions are working as they should. The item-level analysis helps narrow problems to specific questions; sometimes even to a specific part of a question. Further, the quantitative nature of behavior coding data allows researchers to

[2] In relatively rare cases, the behavior coder accompanies the interviewer and conducts the coding on the fly.

identify which questions are problematic and those questions can undergo further revisions and testing prior to the next fielding of the survey.

Willis (2005) states that the key drawback for behavior coding as a survey evaluation method is its reliability. Behavior coding is susceptible to overly subjective opinions on what, for example, is a “major” versus “minor” change in interviewer wording of a question asked. Behavior coders are trained to be precise, but many subjective judgment calls are made by the coders in the actual coding process. One way to combat this limitation is to have coders take copious notes in situations that are unclear or borderline between multiple codes. This provides an opportunity for fixing problems as they are identified after coding.

Important for this paper is precisely this issue of behavior coding reliability. We seek to explore just how reliable six highly trained human coders are when applying these sometimes-subjective codes to the same interviewer and respondent behavior. The way in which behavior coder reliability is typically studied is by using the Kappa statistic (Fliess 1971).

1.2 Kappa and Behavior Coding Reliability

To assess reliability for the behavior coding results in general, we must determine whether the coders were sufficiently trained to apply the same codes to the same observable behaviors. Typically, reliability is assessed using Cohen’s Kappa Statistic.

Kappa, introduced by Cohen in 1960 as an update to Scott’s (1955) pi statistic, is a measure of agreement for categorical level data. It is a better measure of agreement than simply looking at the percent of the time interviewers agree because the Kappa statistic takes into account agreement by chance; i.e., hypothetically, if coders are applying a limited set of codes without even observing an interview, some amount of the time they would agree based on chance and chance alone even though they are just blindly guessing. Kappa subtracts this agreement-by-chance out by approximating only the agreement above and beyond that of chance alone. The Kappa statistic does this arithmetically by taking the actual observed agreement and subtracting the hypothetical agreement that would occur by chance alone. As such, the Kappa statistic typically ranges from 0 to 1, with 1 indicating perfect agreement between coders and 0 indicating agreement equivalent to chance alone (a negative kappa is unlikely but possible, which would indicate agreement below the level of pure chance). While Cohen’s kappa only measures agreement between two coders, Fliess’ kappa (which is based off of Scott’s pi) provides a measure for agreement between more than two coders (Fliess 1971).

The Kappa statistic has important limitations which should be taken into account when interpreting our findings below. The statistic can sometimes provide unexpected results and is inflated slightly by the number of possible codes coders choose from. Further, while there is no universally accepted method of evaluating a kappa statistic, according to Landis and Koch (1977), kappa scores greater than 0.81 indicate an almost perfect level of agreement across coders, 0.61 to 0.80 indicate substantial level of agreement, scores ranging from 0.41 to 0.60 indicate a moderate level of agreement, scores from 0.21-0.40 indicate fair agreement, and scores below 0.20 represent slight to poor agreement.

1.3 Research Objectives

The goal of this paper is to evaluate the reliability of behavior coders with respect to a number of survey research variables. We begin by looking at the reliability of behavior coders for each question in the survey. Next, we look at the overall reliability of coders for coding both interviewer and respondent behavior across the entire survey. We also determine the differing reliability of coders for the English and Spanish language interviews. Previous behavior coding research has shown Spanish coding to be less reliable than English (Goerman et al., 2008). Next, we analyze how reliability differs for different question types (yes/no, open-ended, multiple choice). Another objective is to determine if coders agree more or less depending on the quality of the interviewers' reading and the adequacy of the respondents' answers. Finally, we assess coder reliability as it might be related to length of the questions interviewers read.

2. Methods

This paper is focused not on the quality of questions for any specific study, but rather the reliability of the behavior coders in the evaluation process. This is not a paper evaluating a survey, but an evaluation of behavior coding as an evaluation process. With the above background on behavior coding and the Kappa statistic used to test the coder's reliability in applying codes we will now look at the specific behavior coding scheme being analyzed here.

The taped interviews analyzed here were part of telephone interviews for the 2010 Census Nonresponse Followup (NRFU) survey. This is the survey that collects decennial census data if the Bureau does not have a mail form for a particular household (either because the household never received or did not send back their form through the mail). The survey collects, in twelve questions, basic demographic information. The researchers obtained roughly 200 tape-recorded telephone NRFU interviews. Six bilingual U.S. Census Bureau telephone interviewers were selected based on their speaking and reading fluency in both English and Spanish as well as being reliable interviewers. They received three days of training to behavior code this specific instrument.[3] These six interviewers all coded roughly 30-40 interviews, and all independently coded the same seven interviews for reliability assessment. Five of these seven were done in English and two in Spanish. It is these reliability data that are being analyzed in this study.

Behavior coding for this study was done on both interviewer and respondent behaviors for the first level of interaction; that is, the first speech action of the interviewer was coded as well as the first response by the respondent. If there were multiple levels of exchange–the interviewer and respondent went back and forth more than once–those additional levels of exchange were not coded. Typically, when research intends to identify problem questions, coding the first level of interaction is sufficient because major problems are often evident either when the question is first read or during the initial response from a respondent (Burgess and Paton, 1993; Esposito, Rothgeb, and Campanelli, 1994; Oksenberg et al., 1991; Smiley and Keeley, 1997).

The coders themselves listened to the recording of the interview and followed along with a blank NRFU questionnaire. This allowed the coder to know if the question was read as worded and if the respondent's answer is easily captured by the response options

[3] The training was done in English only.

provided. Coders did not have access to the data that were recorded by interviewers. Their only data source was the audio recordings.

2.1 Codes Used

Specific codes are used to capture ideal and non-ideal behaviors. Codes assigned to interviewer behavior illustrate whether questions are asked as worded. If not, this could indicate that the question is too long, poorly worded or that training is ineffective. If an interviewer skips a question, researchers might suspect the interviewer is judging the information to be redundant, the question to be sensitive or perhaps the skip pattern is too confusing. Interviewer behavior was evaluated using the following codes:

E/S Exact Wording/Slight Change: Interviewer read question exactly as worded or with slight change that did not affect question meaning

MC Major Change in Question Wording: Interviewer made changes to the question that either changed, or possibly could have changed, the meaning of the question

V+ Correct Verification: Interviewer correctly verified information respondent had provided earlier and respondent agrees

V− Incorrect Verification: Interviewer assumed or guessed at information not previously provided (even if correct) or misremembered information when verifying

S Skipped question: Interviewer entirely omitted (answered without reading) an applicable question.

I/U Inaudible/Uncodable: Interviewer was not audible on the tape

Codes assigned to respondent behavior document whether the behavior produced an answer that conforms to what the researcher is attempting to measure. If the wording is awkward or unknown terms are used, the respondent might ask for clarification (Fowler and Cannell, 1996). If the question is too long, respondents might interrupt or ask the interviewer to read the question again. Respondents might even provide an answer that does not conform to the data the question was intended to collect, indicating a cognitive disconnect between respondent understanding and question wording and/or response categories. Respondent behavior was evaluated using the following codes:

AA Adequate Answer: Respondent provided response that can easily be coded into one of the response options
IA Inadequate Answer: Respondent provided a response that cannot easily be coded into one of the response options—often requiring interviewer to probe for more information
QA Qualified or Uncertain Answer: Respondent expressed uncertainty about the response provided or modifies response by placing conditions around their response (e.g., “If you mean this, then the answer is that.”)
CL Clarification: Respondent requested that a concept or entire question be stated more clearly or repeated
DK Don’t Know: Respondent stated they did not have the information
R Refusal: Respondent refused to provide a response

I/U Inaudible/Uncodable: Respondent was not audible

The same seven interviews that the behavior coders all evaluated for reliability purposes each had twelve questions, making the data set a total of 84 data points for interviewer behavior and another 84 for respondent behavior.

3. **Results**

The most general analysis of the reliability of behavior coders for this NRFU test is to look at the Kappa statistic across all questions and all interviews for interviewer and respondent behavior. This is typically what is reported in behavior coding literature about the reliability of a study. The results are found in Figure 1.

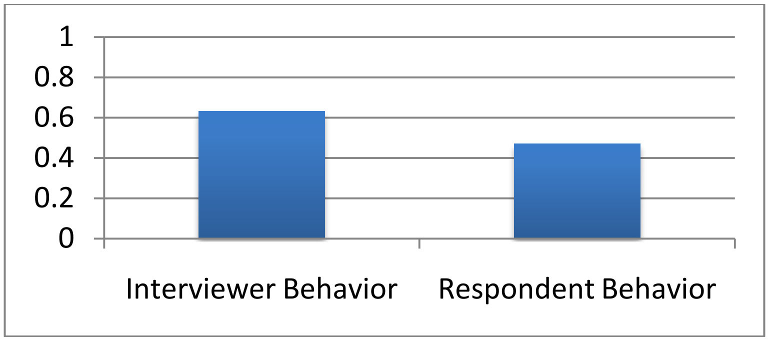

Figure 1. Kappa Scores for Interviewer and Respondent Behavior[4]

The Kappa scores indicate a moderate to substantial amount of agreement between behavior coders, with better agreement around the codes applied to interviewer behavior than to those applied to respondent behavior. Figure 2 shows the same analysis by language, depending on whether the interview was conducted in English or Spanish.

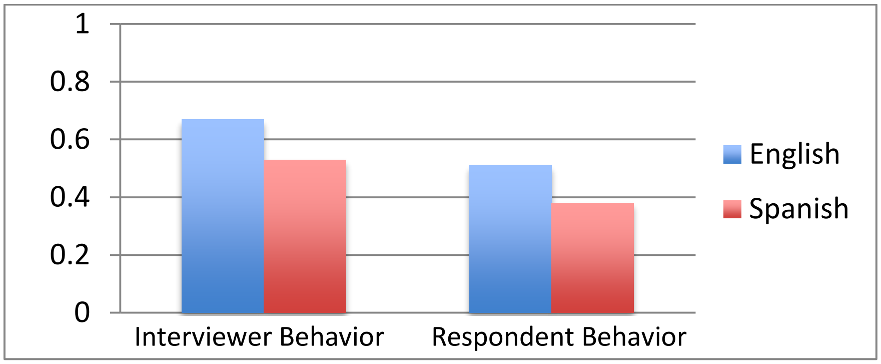

Figure 2. Kappa Scores for Interviewer and Respondent Behavior by Language

[4] 95% Two-tailed confidence intervals: Interviewer behavior = 0.60, 0.69; and for respondent behavior = 0.44 to 0.51. Also, note that since the intervals do not overlap, we can consider the difference between the interviewer and respondent Kappa statistics to be statistically significant at $\alpha<.05$. However, we should also note that statistical significance for the Kappa statistic is often thought to be unreliable and is thus rarely reported.

For both interviewer and respondent behavior coding, agreement was worse for Spanish language interviews, with Spanish respondent behavior having a fair level of agreement. This calls into question the comparability of the English and Spanish language behavior coding results. This also reinforces previous research that has demonstrated Spanish-language coding to be less reliable (Goerman et al. 2008; Edwards et al. 2004). Goerman and colleagues have hypothesized that this may be due to several factors, including that typically questions are written and pretested in English, then translated into Spanish. Additionally, translated instruments may receive less review than English instruments. Finally, the interviewers may have lower fluency in the Spanish language, causing the interactions to be more difficult to understand and to code.

Next, we examine the reliability of coding at a finer level to try to understand why respondent behavior and Spanish-language interviews are less reliably coded. We start this exploration by looking at reliability at the question level. Figure 3 shows question-level reliability. As you can see, it varies dramatically.

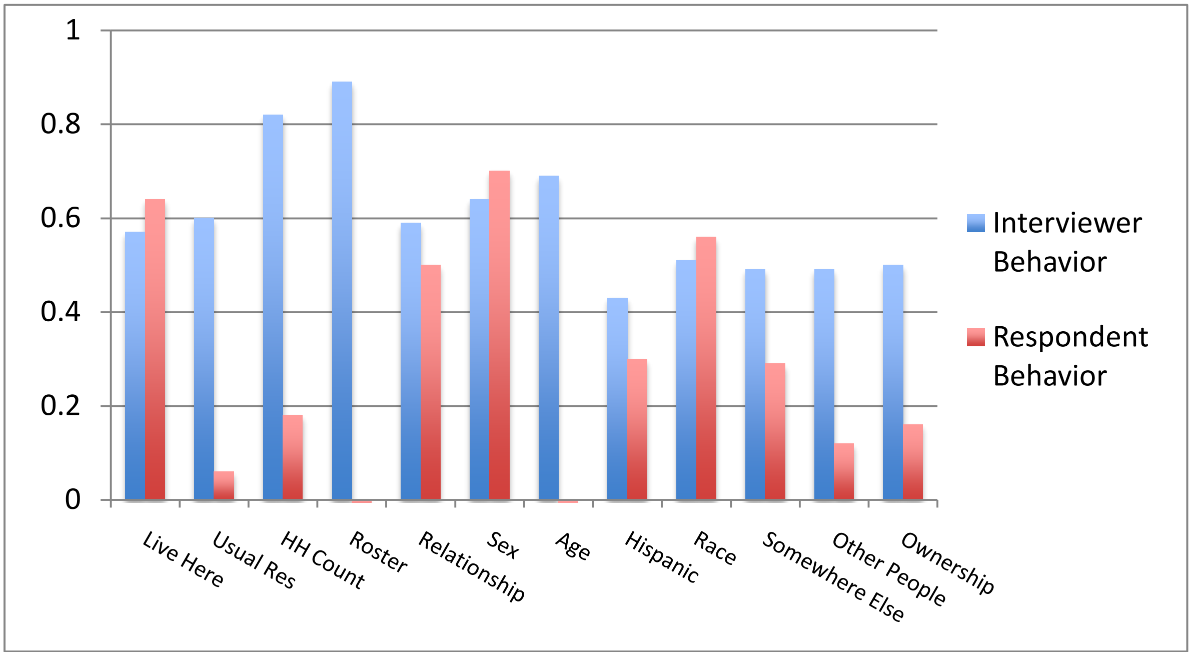

Figure 3. Kappa Scores for Interviewer and Respondent Behavior, by Question Content (English and Spanish)

We suspect that specific question-related issues come into play. The most significant findings here are the extremely low Kappas for respondent behavior for the following questions: Usual Residence, Household Count, Roster, Age and Other People. Reliability for coding interviewer behavior varies as well, but not as dramatically as respondent behavior. We see here that for two of the questions where respondent behavior is the most difficult to code, interviewer behavior is actually the easiest to code (Household count and Roster).

To further explore what might cause these differences, the 12 questions are broken down into three types: yes/no (of which there were two questions), multiple choice (seven questions) and open-ended (three questions). Figure 4 illustrates the reliability for each question type separately for interviewer and respondent behavior.[5]

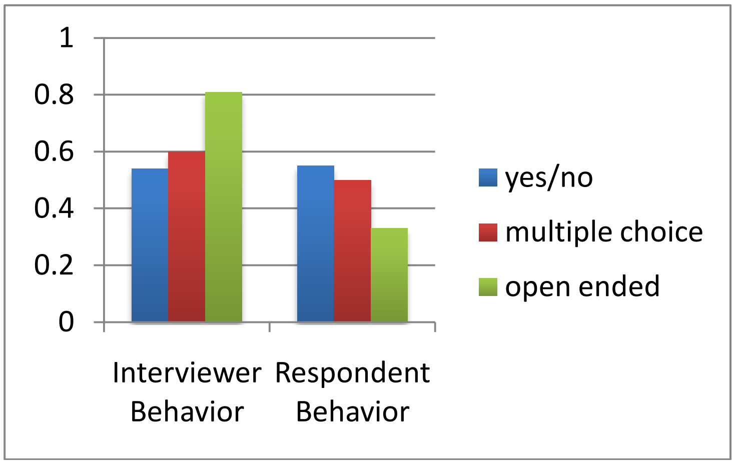

Figure 4. Kappa Scores Interviewer and Respondent Behavior, by Question Type (English and Spanish)

Interestingly, open-ended questions are coded particularly reliably for interviewer behavior and particularly unreliably for respondent behavior. This might be expected since open-ended responses, by definition, leave more ambiguity on just what sort of response the question is looking for. This makes the task more subjective for the behavior coders, and, as previously mentioned, subjectivity is the main source for low inter-coder reliability. We see little difference between yes/no and multiple choice question types.

Our next hypothesis is that it might be easier to code adequate or ideal behavior than variants on that behavior. For example, if interviewers read questions more adequately (that is, an exact reading or minor change), then the coders would be more likely to agree than if interviewers do not. Similarly, coders might also be more reliable when respondents answer adequately. However, Figures 5 and 6 demonstrate the opposite.

[5] Examining the data in Figure 4 by language (data not shown) reveals that the patterns between English and Spanish language interviews are very similar suggesting that this is not a large source of variance for the differences in reliability by language. Thus, those data are not presented separately.

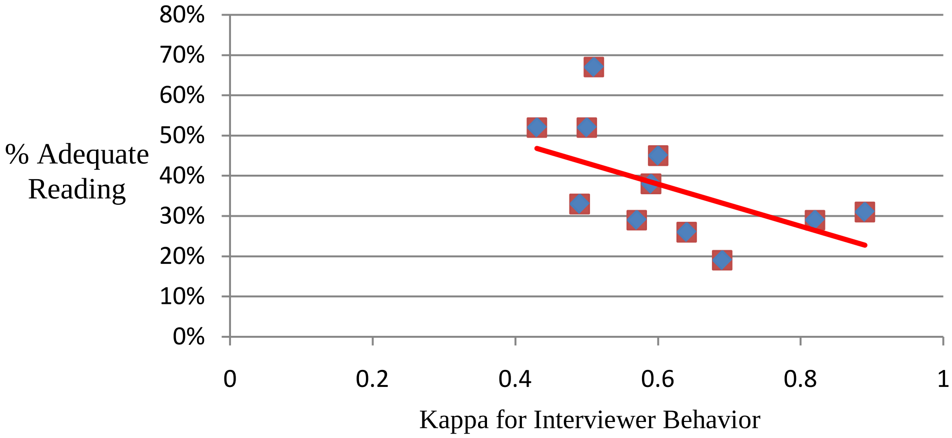

Figure 5. Scatterplot of Kappa Scores for Interviewer Behavior and Percent Adequate Reading for Each Question (English and Spanish)[6]

Figure 5 compares the percent of reportedly adequate question readings by interviewers with behavior coder agreement. Each point in the scatter plot represents a question in the survey. We find an inverse relationship between adequate interviewer reading and the Kappa agreement score of the interviewer's behavior, meaning that questions with more adequate readings tend to produce more disagreement than the inadequate readings. The correlation coefficient is -.54 and is statistically significant at p=.07. What this suggests is that behavior coders agreed more easily when interviewers made mistakes than when the interviewer read the question well. The coders might have more difficulty trying to determine if a change was "minor" or "major" than when the error is much more obvious. This also demonstrates a general point about behavior coding: inter-coder reliability is not a proxy-measure for the survey working well because the coders can all agree that the questions are not being read or responded to correctly.

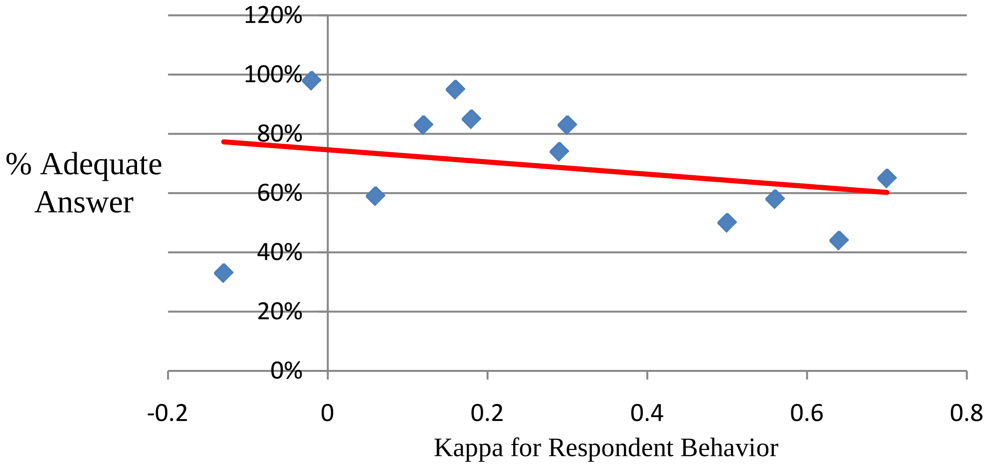

Figure 6. Scatterplot of Kappa Scores for Respondent Behavior and Percent Adequate Respondent Answer for Each Question, with Outlier (English and Spanish)

[6] Note only 11 points in this graphs because two questions, Overcount and H1, both have Kappa scores of .49 and 33% exact reading.

Figure 6 illustrates the same analysis, this time looking at the percent adequate respondent behavior for each question by the corresponding Kappa statistic for respondent behavior. The relationship is, again, negative, however the correlation coefficient is not statistically significant. This means that the negative relationship could be due to chance and not an underlying trend in the data.

However, we ran this analysis again removing the outlier; that is, the question ("age") that has a negative Kappa value. As noted above, sometimes the Kappa statistic produces unexpected results, and for the Age variable we found behavior coder agreement to be less than by chance alone, as is indicated by negative Kappa values. Suspecting this might be an issue with this specific Kappa statistic, we look at this relationship again with the outlier removed. The results are found in Figure 7.

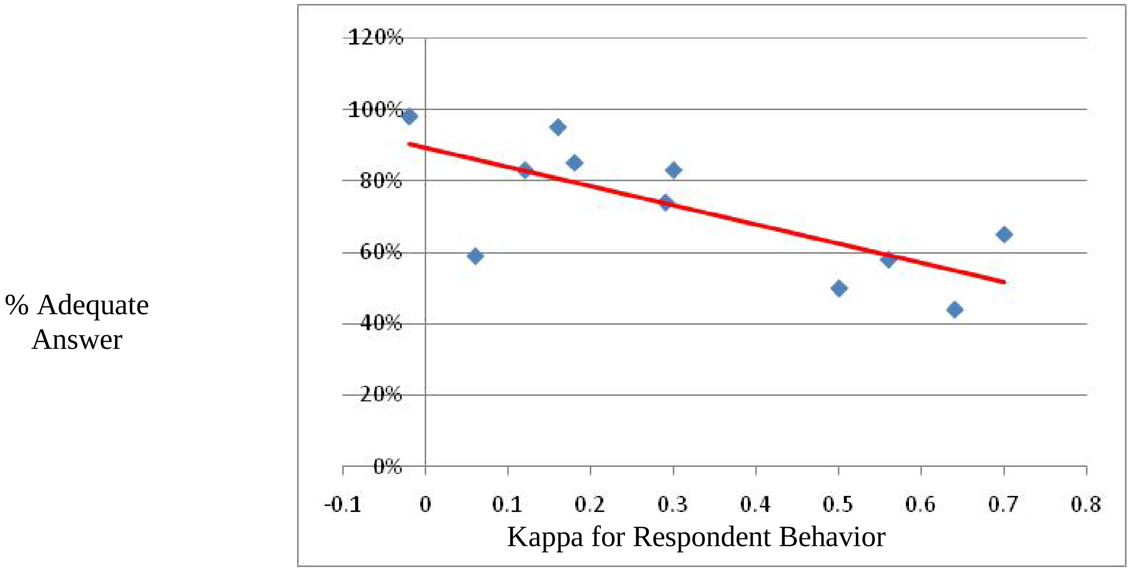

Figure 7. Scatterplot of Kappa Scores for Respondent Behavior and Percent Adequate Respondent Answer for Each Question, Outlier Removed (English and Spanish)

The relationship grows more strongly negative (correlation coefficient = -.66) and is now significantly significant (p=.04). Together with the agreement based on interviewer adequacy chart above, we can make a good case for coders being more reliable when coding problematic interviewer and respondent behavior than when they are adequate.

Given this finding, we looked to see if the accuracy of question reading might be the reason for the differing reliability for English and Spanish behavior coding. We wondered whether English interviews were systematically read and responded to in non-ideal ways, and this might explain the higher Kappa agreement scores when coding English interviews. However, the amount of adequate behavior for both interviewers and respondents is approximately the same for English and Spanish interviews in this study (data not shown separately because of the similarity).

Our final hypothesis was that longer questions (i.e., the more words an interviewer needs to read) would be less reliably coded in both languages. Figure 8 plots each question based on the number of words by the behavior coder agreement for interviewer behavior for that question. As is illustrated by the figure, we found no relationship between question length and coder agreement. Question length does not seem to impact the reliability of behavior coding as an evaluation procedure.

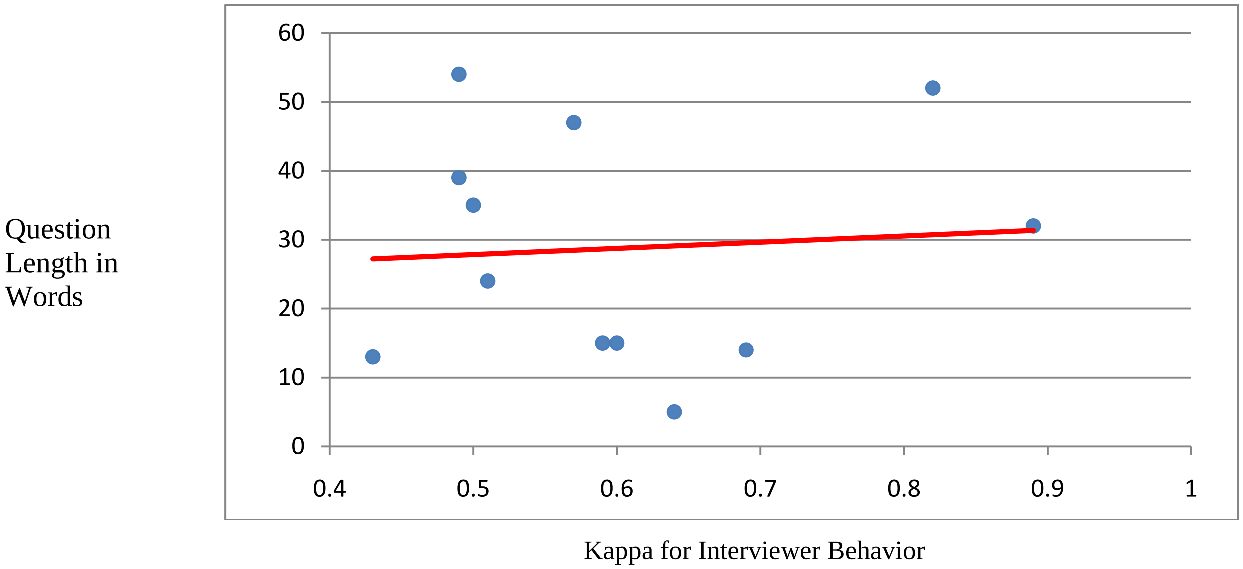

Figure 8. Scatterplot of Kappa Scores for Interviewer Behavior and Question Length in Words (English and Spanish)

4. Discussion and Conclusion

This project is an exploratory study on the relationship between question characteristics and behavior coder agreement. Perhaps the most significant finding is one that reinforces what has already been shown in the literature: that behavior coding of Spanish interviews is less reliable than for English. The significance is that behavior coding, as a method to reduce survey error, is less precise for questionnaires that are translated from English to be executed in Spanish. That coding is less reliable for Spanish interviews means that the comparability of the data may be questioned. Therefore, future research should be conducted on whether behavior coding is a suitable evaluation method for non-English language interviews and, if so, how it can be made more reliable. We also suggest that future research should look at the potential impact of training the coders in Spanish in addition to English.

Next, we also found that the behavior coding of respondent behavior is done less reliably than that for the interviewer. Looking more carefully at question type, we see that coders have specific difficulty coding responses to open-ended questions. Training improvements might be best focused on helping coders more reliably map responses to specific codes in open-ended questions, particularly.

One interesting result was that behavior coders tend to agree more about interviewer and respondent behavior when that behavior is problematic as opposed to adequate. This means that training is doing a good job instructing coders to spot questions and answers that were read or delivered in a problematic fashion, but more attention should be paid to the thin and often subjective line between adequate and inadequate behavior, e.g., just where to draw the line for when an interviewer change to question wording is minor versus major.

In addition to a focus on improving training for problematic areas, this method could be used in real-time during a behavior coding operation to improve reliability. For example, after training, coders could each code a test case of questions. Researchers could look at reliability overall, by language, by question and by question type to find places where coders could use re-training. This would improve the reliability of the behavior coding data produced by the study. In addition, throughout the coding period, the same sequence could be repeated to conduct retraining as necessary to maintain reliable coding.

Finally, this study shows the utility of conducting a more detailed reliability test within a behavior coding study, even post hoc, to allow for recoding when reliability is particularly poor, or to show limitations of the findings. For example, in the current study, the authors may warn the reader that the Spanish language cases were not as reliably coded as the English language cases.

Acknowledgements

The authors would like to acknowledge Laurie Schwede and Matt Clifton for their assistance gathering recordings for this evaluation and Lorraine Randall for preparing the audio tapes for behavior coding. In addition, we thank the Telephone Center Communications Office, particularly John Magruder and Olga Koly, as well as the staff from the Tucson Telephone Center for the manual behavior coding work. We appreciate comments from our critical reviewers as well, Heather Madray, Sarah Heimel, Ashley Landreth, Dave Sheppard, Hilda Dimmock, Geoff Jackson, Florence Abramson and Nancy Bates. And finally, we thank Katie Drom for her assistance with the qualitative analysis and Sarah Wilson and Michelle Smirnova for their editorial assistance and fact-checking.

References

Burgess, M. and D. Paton, (1993). Coding of Respondent Behaviours by Interviewers to Test Questionnaire Wording. Proceedings of the Survey Methods Research Section. Alexandria, VA: American Statistical Association.

Cannell, C., F. Fowler, and K. Marquis, (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: US Government Printing Office.

Edwards, S. , Zahnd, E. , Willis, G. , Grant, D. , Lordi, N. and Fry, S. (2004). "Behavior Coding across Multiple Languages: The 2003 California Health Interview Survey as a Case Study" Paper presented at the annual meeting of the American Association for Public Opinion Research, Pointe Hilton Tapatio Cliffs, Phoenix, Arizona.

Esposito, J., J. Rothgeb, and P. Campanelli, (1994). The Utility and Flexibility of Behavior Coding as a Methodology for Evaluating Questionnaires. Paper presented at the American Association for Public Opinion Research Annual Meeting, Danvers, MA.

Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382

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

Goerman, P., J.H. Childs, and M. Clifton, (2008). Explaining Differences in Inter-coder Reliability between English and Spanish Language Behavior Coding Research. Proceedings from the 2008 Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association.

Landis, J. R. and G.G. Koch, (1977) "The measurement of observer agreement for categorical data" in *Biometrics*. Vol. 33, pp. 159–174

Oksenberg, L., C. Cannell, and G. Kalton, (1991). New Strategies for Pretesting Survey Questions. *Journal of Official Statistics* 7 (3): 349-394.

Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding." Public Opinion Quarterly, 19(3), 321-325.

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

Smiley, R. and C. Keeley, (1997). Behavior Coding of the ICM Computer Assisted Personal Interview. 1996 Community Census Results Memorandum Number 23. U.S. Census Bureau, Washington, DC.

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