Measuring Data Quality for Telephone Interviews

Ryan King
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
Past research for telephone interviews has focused mostly on service quality, which evaluates the interviewers' actions during the interview. To measure the quality of the data, the simple thing to do is look for a keying error by interviewers. However, in interviewing situations it's not always clear what "truth" is, and therefore one must also consider other aspects of the interview process to determine if the correct keying action was taken. This includes how the interviewer asks the question and how the respondent answers, or reacts, to the question. If any part of this process has an interaction different than what is expected, it may affect the quality of the data collected. In the 2010 Census we plan that over 2.5 million telephone interviews will be conducted, and thus the quality of the data have far reaching impacts. To develop a method for use in the 2010 Census to measure data quality, the Census Bureau used taped telephone interviews from the 2006 Coverage Followup, an interview designed to correct the household roster on previously submitted Census questionnaires. This paper will discuss how we developed that method, how it was implemented, and some of the results we obtained.

Key Words: data quality, keying error, telephone interview

1. Background[1]

Past research for telephone interviews has focused mostly on service quality, which evaluates the interviewers' actions during the interview. To date there have been relatively few studies of the quality of the data coming out of computer assisted telephone interviews (CATI). A study by Kennedy, Lengacher, and Demearth that studied keying errors by monitoring a CATI, saw an error rate of 0.6 percent (Lepowski, 1995). In another study done by Lepowski, Sadosky, and Weiss that studied keying errors by using taped interviews, saw an error rate of 0.1 percent (Lepowski, 1995). As you can see from these past studies, the error rates found are quite small.

To measure the quality of the data, the simple thing to do is look for a keying error by interviewers. However, in interviewing situations it's not always clear what "truth" is, and therefore one must also consider other aspects of the interview process to determine if the correct keying action was taken. This can include the type of interview mode, the type of question, and the interviewer and respondent interaction (Lepowski, 1998). If any of those aspects creates an interaction different than what is expected, it may affect the quality of the data collected.

In the 2010 Census we plan to conduct 2.5 million telephone interviews, and thus the quality of the data have far reaching impacts. To develop a method for use in the 2010 Census to measure data

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

quality, the Census Bureau used taped telephone interviews from the 2006 Coverage Followup (CFU), an interview designed to correct the household roster on previously submitted Census questionnaires.

2. Limitations

In order for an interviewer to record the interview, the respondent needed to give their verbal consent that it was okay with them. Thus both the interviewer and the respondent knew they were being recorded, which may have modified their behavior. In addition, both interviewers and respondents were not selected randomly. Each interviewer was asked to record at least one to three English interviews, of which the respondents were a sample of convenience.[2]

The results presented here come from a relatively small sample of interviews. In total there were 122 interviews available, but only 103 of these interviews could be used. Some reasons that a tape could not be used include the following: coders did not hear the respondents consent to be taped at the beginning of the tape (as mentioned above), the tape was mislabeled (either the wrong case or the tape indicated it was in English but was Spanish on the tape), or coders indicated the tape was inaudible.

There were some questions asked within the interview that we were not able to evaluate, because they were not asked often enough. For example, the survey collects demographic information about the persons in the household, but only if the data was missing from the original return, or a person was added to the roster during the interview. A relatively small number of these data items are missing on the initial return, and only a few cases had persons added to the roster, so the questions asked by interviewers to obtain this information were not asked often.

3. Methods and Results

The purpose of this study was two-fold:

1. To determine a methodology that could be implemented by coders during the 2010 Census with little survey experience, a few days of training, and usable for a large number of cases.
2. Using that methodology, determine an expected data quality rate.

While one could simply listen to the tapes to hear the respondents answer to each question, and then look at the output data to determine if the correct response was collected, we believed there was more to collecting the correct response. We believed that the verbal interaction between the interviewer and the respondent would also play a vital role in whether the correct answer was obtained. Lepkowski and his colleagues influenced this thinking; in that they also attempted to account for various interviewer and respondent interactions when coding outcomes of the quality of data coming out of CATI interviews (Lepkowski, 1998).[3] Thinking like this combines aspects of service quality monitoring and simple methods of identifying keying errors through the use of audiotapes into account.

[2] Interviewers, who were bilingual in English and Spanish, were asked to attempt to record one interview in English, one interview in Spanish, and if possible a third (and fourth) interview in Spanish, but if that was not possible then an English interview was acceptable. Interviewers who were not bilingual in Spanish were asked to record three interviews in English.

[3] The paper also looked at how that quality compared to paper and pencil telephone interviews.

Therefore we developed a coding scheme that would code each item (or question) as 'Accurate', 'Inaccurate', or 'Uncertain' based on the verbal interaction between the interviewer and the respondent, the respondent's response, and the value that was recorded in the output. The rules to be applied to each code were as follows:

- If the interviewer read the question verbatim, there was a simple exchange between the interviewer and the respondent, and the response matched the output, then the question would be coded as 'Accurate.'
    - A simple exchange is defined as one where the respondent gives a question appropriate answer that the interviewer can input into the instrument without further probing.
- If the interviewer did not read the question or if the respondent's response did not match the output, then the question would be coded as 'Inaccurate.'
- If the question was read to the respondent, but not verbatim, if the question had a more complex exchange between the interviewer and the respondent, or if the audio from either the interviewer or respondent was inaudible, then the question would be coded as 'Uncertain.'
    - A complex exchange is defined as one where a "conversation" occurs between the interviewer and the respondent, whether it is about the question or not.
    - We asked coders to use this code, even if the coder believed the interviewer got the answer correct. The reasoning behind this was that we could not decide how far off script should be allowed without the intent of the question being changed.
    - The coders were also instructed to take detailed notes, so that we could later learn what types of interactions were causing these 'Uncertain' codes, and reassess how to code some of these situations.

The following example shows how the same question could have received different codes:

- Accurate
  Interviewer: In the spring of 2006, was anyone attending college?
  Respondent: Yes.
  Output: Value of 'Yes'
- Inaccurate
  Interviewer: In the spring of 2006, was anyone attending college?
  Respondent: Yes.
  Output: Value of 'No'
- Uncertain
  Interviewer: Was anyone attending college?
  Respondent: Yes.
  Output: Value of 'Yes'

To determine if this coding scheme would prove useful, we distributed a subset of the tapes to the coders and asked them to listen to and code ten of the interviews.[4]

There were four persons chosen to do the coding of the tapes for this survey. The four persons had varying degrees of time of employment at the Census Bureau, as well as varying degrees of

[4] One of the coders coded eleven interviews.

knowledge about the survey they would be coding. In addition, there was a brief training session that provided an overview of the survey and its purpose, and on the instrument that would be used to record the codes (an Excel spreadsheet). These were seen as positive factors, since these features emulated some of the features that the coders in 2010 may have.

Of the 41 tapes that were listened to, only 31 were actually coded. Reasons the coders gave as to why they could not be coded were that the tape was inaudible, or the tape had been mislabelled (for example, the interview was in Spanish or the wrong household was on the tape). Coders did indicate that the coding process was fairly simple to implement after one to two interviews. Most of the items coded had at least a 90.0 percent or greater accurate rating.

After implementing this coding schema on the subset of tapes the team decided to move forward and code the remainder of the tapes in this fashion. The reason being that after reviewing the 'Uncertains' the team could not determine how to provide rules such that they could be classified as 'Accurate' or 'Inaccurate.'

To get a measure of the number of items that were coded as 'Accurate', 'Inaccurate', or 'Uncertain,' we counted up the number of each of these by case. To get a total number of items with codes, we summed up the counts for 'Accurate', 'Inaccurate', and 'Uncertain.' We then calculated the proportion that each code received of that total number, and averaged that number across all cases. Thus the numbers in Table 1 (as well as Table's 2 and 3) represent the average percentage of items for all cases that received a certain code. As seen in Table 1 below, we saw that when looking at all of the items asked in the interview, that about 97.10 percent of the time items were coded as 'Accurate', about 0.30 percent of the items were coded as 'Inaccurate', and about 2.60 percent of the time items were coded as 'Uncertain'.

Table 1: Type of Code and Percentage of Items that Received that Code: Initial Second Round of Coding

Type of Code	Percent (Standard Error)
Items that received an 'Accurate' code	97.10 (0.61)
Items that received an 'Inaccurate' code	0.30 (0.16)
Items that received an 'Uncertain' code	2.60 (0.51)

To test how consistently the coders were applying the codes as intended, during this second round of coding we had each of the coders code the same two interviews. To do this we calculated a Kappa statistic to measure the inter-coder reliability. The Kappa statistic provides a conservative measure of agreement among coders in their application of behavior codes, because it accounts for the possibility of agreement by chance (Fleiss, 1981). According to Fleiss, Kappa scores greater than 0.75 indicate an excellent level of agreement across coders, while scores ranging from 0.40 to

0.75 indicate a good to fair level of agreement; scores below 0.40 represent poor agreement (Fleiss, 1981). The Kappa statistic achieved from these codings was 0.00, which indicates that there is very poor agreement.[5] However, when looking at the raw data we see that generally our coders were producing the same results. For 48 of the 60 items, or 80.0 percent of the items that were coded, the same code was produced by all coders. Thus because our coders did not have much variation in the codes they applied, nor did all coders use all possible codes, is a possible reason such a low Kappa statistic was produced. For the few items that caused a majority of the disagreement, the group discussed these items, and were able to come to an agreement about what the correct codes should have been.

After discussing the initial results from the second round of coding, coders indicated that there were some questions that were coded as uncertain that could be coded as accurate or inaccurate if the rules were modified to allow for some more complex interactions between the interviewer and the respondent. Thus the following rule was appended to the 'Accurate' and 'Inaccurate' codes:

- When a conversation occurs between an interviewer and a respondent and the interviewer provides information about how to answer the question from either training materials or help materials and the respondent provides a final definitive answer, then the question would be coded as 'Accurate' or 'Inaccurate,' as appropriate.

After modifying the rules, the number of responses that received an 'Accurate' code increased. For example, one question, the review roster question, went from 81.4 percent 'Accurate' to 93.1 percent 'Accurate.'

Table 2 shows the results based on the modifications to the coding rules at the item level. As expected, the rate at which items received 'Uncertain' codes went down. Also of note is that the rate at which items are coded to be 'Inaccurate' is still similar to past studies, even though it went up slightly.

Table 2: Type of Code and Percentage of Items that Received that Code: Updated Second Round of Coding

Type of Measure	Percent (Standard Error)
Items that received an accurate code	98.66 (0.42)
Items that received an inaccurate code	0.45 (0.19)
Items that received an uncertain code	0.89 (0.29)

[5] The above Kappa statistic was produced using SAS procedure that calculates a Kappa for each pair of coders, and then aggregates it to an overall Kappa statistic. We had also calculated a Kappa outlined in Fleiss' book, Statistical Methods for Rates and Proportions (2003). This method produced a Kappa score by rating category, across all questions, and then produced an aggregate Kappa score from those. The Kappa produced was 0.02. The difference in the two scores is not practically significant, and we conclude that the slightly different methods of producing the aggregate Kappa is what is causing the small discrepancy.

When planning the 2010 Census, it was determined that it would be too costly and too time consuming to code every variable in the CFU interview. Thus a subset of questions, deemed "Critical Items," from all questions in the interview were chosen to be coded. Table 3 shows the accuracy and error rates that are calculated when only looking that the items that will be coded during the 2010 Census. 'Uncertain' codes are excluded because it was believed those outcomes should not count against the final accuracy rate, since we could not say with certainty whether they were 'Accurate' or 'Inaccurate.' As you can see the percentage of items receiving 'Accurate' codes is quite high at 99.64 percent and the percentage of items receiving 'Inaccurate' codes, is again similar with past studies, with it being 0.36 percent.

Table 3: Type of Code and Percentage of Items that Received that Code: Only Critical Items That Will Be Looked at in 2010

Type of Measure	Percent (Standard Error)
Items that received an accurate code	99.64 (0.21)
Items that received an inaccurate code	0.36 (0.21)
Items that received an uncertain code	--- (---)

4. Conclusion

Like previous research, we saw only a small amount of entries, about 0.45 percent, with a coding of 'Inaccurate.' Subsequently, we also only saw about 0.89 percent of all entries receive an 'Uncertain' coding. With more intense training, some of these 'Uncertain' codes could possibly be changed to 'Accurate' or 'Inaccurate' codes. In the end, we are confident that nearly all entries are input accurately.

For the 2010 Census, the Census Bureau will be selecting a sample of CFU cases to code using the final methodology described within this document. However, due to budget constraints, we will not be able to code the entire interview, or even all items that we consider critical to the CFU interview. Instead, only a subset of items that are considered critical to the CFU interview will be coded. Overall we expect to see an item accuracy rate in the range of 94.0 to 99.0 percent. This is the first time the Census Bureau has done something like this on this survey, so we hope to learn a lot about how well the method works. Thus, any lessons learned from using this methodology during the 2010 Census could then be used to improve it where applicable. We can then apply a similar method to inter-decennial tests of this survey, and ultimately, again during the 2020 Census.

Acknowledgements

I would like to thank the people who aided in the research and development of this methodology: Elizabeth Krejsa, Dave Sheppard, Sarah Brady, Conray Boyd, Dana Cope, Frank Anderson, and Aref Dajani.

References

Davis, Diana; Allen, Samuel; "Cognitive Testing and Evaluation of Coverage Issues and Residency Rules: Project C, Behavior Coding Analysis of Coverage Followup (CFU) Telephone Interviews, 2006 Census Test," Development Associates; March 29, 2007.

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

Fleiss, J. (2003) Statistical Methods for Rates and Proportions. John Wiley & Sons, New York.

Landreth, Ashley; Krejsa, Elizabeth; Karl, Leann; "Behavior Coding Analysis Report: Evaluating the Coverage Research Follow-Up (CRFU) Survey for the 2004 Census Test Administered Using Telephone and Personal Visit Survey Modes," U.S. Census Bureau; July 18, 2005.

Lepkowski, James, Sadosky, Sally, and Weiss, Paul (1998); "Mode, Behavior, and Data Recording Error," *Computer Assisted Survey Information Collection*, John Wiley & Sons, Inc., 1998.

Lepkowski, James; Sadosky, Sally; Couper, Mick; Chardoul, Stephanie; Carn, Lisa; Scott, Lesli Jo (1995); "A Comparison of Recording Errors Between CATI and Paper-and-Pencil Data Collection Modes," Survey Research Center, Institute for Social Research, University of Michigan.