Behavior Coding Assessment of the Quality of Real-time Telephone Survey Interpreters

Michael W. Link[1*], Larry Osborn[2], Marta Induni,[3] Michael P. Battaglia[2], and Martin R. Frankel[4]

[1]Centers for Disease Control and Prevention, [2] Abt Associates Inc., [3] Public Health Institute, [4] Abt Associates and Baruch College, CUNY; *Contact author: 4770 Buford Highway NE, MS: K-66, Atlanta, GA, 30341-3717, MLink@cdc.gov, 770-488-5444

Abstract

Real-time interpretation during a survey can expand the number of languages in which surveys are offered. There are questions, however, about the quality of the interpretation process given that the interview is typically not pre-translated. A detailed assessment of the quality of this approach is provided using behavior coding of interviews conducted with respondents who otherwise would have been finalized as "language barrier nonrespondents." Interviews were recorded and behavior coded, quantifying for each question (1) the accuracy of the question interpretation, (2) the accuracy of the interpreted response, (3) the degree of difficulty administering the question, (4) the number of times the question needed to be repeated, and (5) the number of times the interpreter and respondent engaged in dialogue that was not relayed to the interviewer. The approach produced favorable results, with less than a 4% error rate for interpretation of the questions and a 1.4% error rate in interpretation of survey responses.

1. Introduction

Despite the dramatic growth in the non-English/non Spanish-speaking population in the United States over the past decade, most surveys are still conducted only in English or Spanish. One approach to expanding the number of languages offered in telephone surveys is to use a third-party interpreter, offering real-time interpretations on the telephone as the survey is conducted. This approach allows the interview to be conducted in a broader range of languages and typically makes more effective use of language specialists than does hiring native speakers as interviewers. While research has shown this to be an effective means of improving survey response among non-English/Spanish speakers, there are still questions about the quality of the interpretation process given that, for most languages, the interview is not pre-translated. This raises concerns about how real-time interpretation might affect survey responses (Murray, Battaglia, and Cardoni 2004).

We provide the first detailed assessment of the quality of the real-time interpreter approach, using behavior coding of interviews conducted as part of the 2005 California Behavioral Risk Factor Surveillance System (BRFSS). The interviews were conducted with a set of respondents that otherwise would have been given a final code of "language barrier" and counted as survey nonrespondents. With the permission of the respondents, interviews were recorded and later behavior coded, quantifying for each question administered (1) the accuracy of the question interpretation, (2) the accuracy of the interpreted response, (3) the degree of difficulty administering the question, (4) the number of times the question needed to be repeated, and (5) the number of times the interpreter and respondent engaged in dialogue that was not relayed to the interviewer.

2. Methods

The assessment was conducted in California in three phases: (1) identification of eligible telephone numbers, (2) contacting and interviewing sample members using a real-time interpreter, and (3) behavior coding of the recorded interviews. Telephone numbers were sub-sampled from the regular, monthly BRFSS sample records if the case was finalized as a "language barrier problem," meaning that no one in the household who spoke English or Spanish (the two languages in which BRFSS is conducted) could be reached. To increase the potential sample size for the pilot study, cases initially sampled from January 2005 onward were considered eligible for inclusion, even though interviewing did not begin until September 2005. Telephone numbers sampled from January through August were considered "retrospective" cases for the purposes of the analysis presented here because contacts via the interpreter were initiated a month or more after the last BRFSS call attempt. Those numbers sampled September through December were considered "concurrent" cases because contact attempts via the interpreter with these cases were generally initiated within a month of the last BRFSS call attempt.

Contacting and interviewing the language follow-up cases began 1 September 2005. Interviewers from the same survey research group that conducts the California BRFSS initially called all retrospective telephone numbers to determine as best as possible the language spoken in the household. This was accomplished by contacting the households

and attempting to obtain an answer to one of the following questions: 1) What language do you speak? or 2) What country are you from? Researchers anticipated that many non-English speakers would still be able to answer one of these questions, even when posed in English. To initiate follow-up contact, interviewers first contacted the interpreter service to obtain the assistance of an interpreter fluent in the language thought to be spoken in the household. It is important to note that each of the interpreters on the project were bound by the same confidentiality rules and requirements as the interviewers. Next a three-way call was established between the interviewer, the interpreter, and the sampled telephone number. In situations where the language could not be determined during the initial contact, or was different than initially indicated, a language specialist at the interpreter service came onto the line to assist the interviewer in identifying the correct language and accessing an appropriate interpreter. For concurrent cases, the BRFSS interviewers asked the same questions and entered the likely language into case notes for use by the language follow-up survey interviewers. Contacting and interviewing for numbers resulting in an eligible household then proceeded with the interviewer administering the survey and the language specialist providing interpretation of the question for the sample member and the response for the interviewer. With permission from the sample member and interpreter, the interviews were recorded for later processing.

Because the fielding period extended over two calendar years, the 2006 BRFSS core questionnaire was used for all interviews. This was logistically easier and less expensive than using different questionnaires for different years. Further details on BRFSS survey design, methodology, and questionnaire are available elsewhere (Mokdad, Stroup, and Giles 2003 and at http://www.cdc.gov/brfss).

After completing the telephone interviews, the taped interviews were behavior coded by trained language specialists from an outside language service vendor (not the vendor providing the interview interpreters). The coders were fluent in the language in which the interview was conducted and were trained on the procedures for behavior coding. The coders assessed administration of each question on the following five attributes:

- Was the question interpreted accurately? (1 = least accurate, 4 = most accurate)
- Was the response relayed accurately? (1 = least accurate, 4 = most accurate)
- Were there concepts in the question that appeared to be difficult to interpret accurately? (yes/no)

- How many times did the question need to be repeated to the sample member? (number of times)
- Were there side conversations between the interpreter and the respondent that were not interpreted for the interviewer? (yes/no)

The final dataset contained the following information: (1) call history information from the original BRFSS survey; (2) call history and case disposition information from the language follow-up; (3) questionnaire responses for the completed questionnaires; and (4) the behavior coding assessments for each question. The analysis focused on three key areas: final case dispositions, demographic characteristics, and question-level and interview-level assessments of the quality of the interpretations.

First, we examined response rates and final case outcomes for the language follow-up. Response rates were calculated using response rate formula #4 recommended by the American Association for Public Opinion Research (AAPOR 2004).

Second, demographic characteristics of the respondents were compared between the 2005 California BRFSS and the language follow-up. To adjust for survey design, both sets of cases were weighted inversely by the number of landline telephones in the household and the number of adults in the household.

Third, we assessed the quality of interpretations at both the question level and the interview level using the five behavior coding attributes. Each attribute was recoded to form five dichotomous quality indicators. Questions that were graded as 1 or 2 for the interpretation accuracy and response accuracy measures were respectively coded 1, indicating there was a problem with interpretation of that question. Questions where the coder indicated that there were concepts that were difficult to interpret accurately and where there were side conversations between the interpreter and respondent were each coded as 1. Finally, if the question was repeated one or more times, the question was coded as 1 in terms of the need for the question to be repeated.

At the question level, we calculated for each of the five dichotomous indicators the proportion of responses to each question where an error or problem with interpretation was noted. In effect, these measures showed the number of times we found an error or problem with interpretation or question administration for every 100 times the question was administered. To determine if the errors or problems identified were related to the types of questions being asked, we compared these measures in terms of

position of the question in the questionnaire (first third, middle third, or final third), if the question was a primary question (or "gate" question) asked of everyone or a follow-up question asked only of a subset of those responding to the primary question, the type of response options provided in the question (yes/no, categorical, numeric, or Likert scale), and the subject area content of the questions. Significance was assessed by comparing proportions, using an F-test of means.

To assess quality at the interview level, we calculated the proportion of questions in each interview where an error or problem was indicated. This was calculated for each of the five measures individually, thereby giving us interview-level scores for each measure. We then compared how these measures varied across different types of respondents (in terms of sex, age, education, family income, and language spoken). Significance was assessed by comparing proportions, using an F-test of means.

3. Results

The preliminary dataset included 988 cases finalized in the California BRFSS as having a language barrier and treated as nonresponses. Completed follow-up interviews were obtained from 201 of these cases, of which 195 had error-free tapes that were behavior coded. Assessment of participation rates are based on the entire set of cases, while subsequent analyses of demographics and interpretation quality are limited to the cases for which both a complete interview and complete behavior coding were available.

3.1 Participation rates

Overall, the response rate for the language pilot was 28.9%. There were no significant differences in the response rates between the retrospective and concurrent cases, so these data were combined for subsequent analyses.

The response rate for the 2005 California BRFSS was 27.7%. Using the final case distribution for the language follow-up cases, we estimated that if the interpreter approach had been used throughout the year, the 2005 California BRFSS response rate would have increased by less than 1 per cent, to 28.5%.

3.2 Demographic characteristics of respondents

Among the initial 195 language follow-up respondents, interviews were conducted in 26 different languages (Table 1). Asian languages predominated, with Vietnamese, Mandarin, Cantonese, and Korean being the most prominent, comprising more than 60 per cent of the interviews. Interviews were conducted in Russian approximately 10% of the time.

Respondents to the language follow-up differed significantly from the larger set of California BRFSS respondents in several demographic characteristics (Table 2). Nearly three-fourths (72.3%) of respondents were Asian, compared with 6.1 per cent in the BRFSS. Language follow-up respondents were also more likely to be aged 65 or older (37.7% versus 15.2%), to have a high school or less education (63.0% versus 40.4%), and to have a household income below $25,000 (62.2% versus 30.9%).

3.3 Quality of Interpretation

We examined the quality of the interpretation and respondent-interpreter-interviewer interaction at both the question and interview levels. First, focusing on the question-level assessment, problems were found with the initial interpretation 3.6% of the time a question was administered. Problems were most prevalent for the questions, "One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. During the past 30 days, on the days when you drank, about how many drinks did you drink on average?" (12.0%), "How often do you get the social and emotional support you need?" (11.8%), "Has a doctor, nurse, or other health professional ever told you that you had angina or coronary heart disease?" (10.8%), and "Considering all types of alcoholic beverages, how many times during the past 30 days did you have (five or more [males] / four or more [females]) drinks on one occasion?" (10.5%). No interpretation errors were reported for 2 of the 75 questions.

Errors in relaying the response back to the interviewer were less prevalent, occurring, on average 1.4 per cent of the time. Error rates were highest for questions on race ("Which one of these groups would you say best represents your race?" [asked of those who indicated in an initial question that they were of multiple races]) (7.0%) and binge drinking during the past 30 days (6.6%). For nearly 30 per cent of the questions (22 of 75), there were no problems found in relaying the response.

Conceptual problems making interpretation difficult occurred, on average, 4.0 per cent of the time. This was most problematic for the question on social and emotional support (16.6%) and angina and cardiovascular disease (13.8%), as well as a question asking, "About how long has it been since you last visited a doctor for a routine checkup?" (10.4%). Only the question on pregnancy status had no reported conceptual problems.

Questions needed to be repeated, on average, 10.7 per cent of the time, with repeat rates of 10 per cent or higher for 30 of the 75 questions. Among the most often repeated were the questions on

social and emotional support (34.3%), physical health ("Now thinking about your physical health, which includes physical illness or injury, for how many days during the past 30 days was your physical health not good?") (27.8%), and last routine checkup (27.2%). There were no questions that did not have to be repeated for at least one respondent.

Side conversations between the interpreter and the respondent that were not interpreted for the interviewer occurred 4.4 per cent of the time. Side conversations were most likely to occur during administration of the questions on number of days of poor physical health (15.5%), social and emotional support (14.1%), and last routine checkup (14.1%). Side conversations did not occur during administration of 7 of the 75 questions.

We also assessed the degree to which question placement within the questionnaire, the type of question, and the response format might have influenced the proportions on these five quality measures (Table 3). Errors in interpreting questions for the respondent did not appear to be significantly related to question placement, question type, or response format. Errors in relaying responses to the interviewer were higher for questions with categorical responses (such as type of employment or marital status) than they were for other types of response formats. Questions towards the end of the survey were less likely to experience errors than were those administered during the first two-thirds of the survey. Primary questions, that is, those asked of all respondent or those that serve as "gate questions" to a set of follow-up questions were more likely to have concepts that were difficult to interpret than were follow-up questions (5.4% versus 2.4%). Questions with a "yes/no" format were less likely than other response formats to require that the question be repeated and less likely to generate side conversations between the interpreter and the respondent. Conversely, primary questions were more likely than follow-up questions to stimulate side conversations.

Next we examined the quality of the interpretation looking at the interview-level. For each of the five measures, we calculated the proportion of questions in an interview that resulted in errors or problems for each respondent. These interview-level proportions were then compared across a number of different demographic groups (Table 4). The language spoken had the greatest impact, with significant variation found on all five measures. Respondents speaking less prominent languages (that is, a language other than the five most prevalent languages) were more likely to have interviews with errors in interpretation of the question and of the response. They also had a significantly higher

proportion of questions where conceptual issues made it more difficult to interpret the question. Russian-speaking respondents were the group most likely to have questions repeated and to engage in side conversations with the interpreter. Among the other demographic characteristics examined, women were more likely than men to have questions repeated. Respondents aged 70 or older were more likely than younger respondents to engage in side conversations with the interpreter. Neither education nor income was significantly related to any of the five quality measures.

4. Discussion

Since most health surveys of the general population in the United States are conducted in English, or in some cases Spanish, people who speak a language other than these two predominant languages are often under-represented or not represented at all by these surveys. As a result, the health risks and problems that they face may be inadequately described in public health statistics. Although the percentage of such individuals may still be modest on a national scale, non-English/Spanish speakers often comprise a significant percentage of the population in local areas. In turn, these pockets of individuals may have significant health care needs, which could go undocumented in many health surveillance efforts.

Real-time interpretation may be an effective technique for including these individuals in survey research efforts. We were able to complete interviews with just over 25% of the cases initially finalized in the California BRFSS as "nonrespondent, language barrier" cases. These individuals would have been classified as nonrespondents due to a survey design limitation (not offering the survey in the respondent's language) rather than to their unwillingness to participate. Moreover, these individuals had a very different demographic profile than those typically interviewed as part of the BRFSS in California. The approach reduced the potential for nonresponse bias, therefore, by reducing the level of nonresponse and improving participation among respondents with characteristics different from the original pool of respondents.

In terms of quality assessment, overall percentage of error in administering the questions appears modest at less than 4%, whereas error in interpretation of the responses was much lower at just over 1%. These error rates were even lower for the more prevalent languages (Vietnamese, Mandarin, Cantonese, Korean, and Russian). Error rates varied considerably by question type, placement, content, and response format. Questions that required repeating or which stimulated additional side

conversation between the sample member and the interpreter (without including the interviewer) also tended to be those that had apparent conceptual difficulties, making interpretation more problematic and requiring additional explanation. Cognitive testing and interpreter training could improve the process in both of these areas. Currently the BRFSS questions are cognitively tested only in English. Researchers need, however, to be cognizant of the customs, values, and beliefs of persons in linguistically and culturally diverse communities, particularly because they relate to the sharing of personal information, including health-care practices and health conditions (Hilton and Skrutkowski 2002). Focus groups and cognitive interviews of people from various backgrounds can help determine whether respondents will interpret and respond to survey requests and questions as intended (Eyton and Neuwirth 1984). Likewise, additional training of the interpreters stressing the importance of their remaining neutral third-party facilitators in the interview process could help to reduce the number of side conversations in which the interviewer is excluded.

The financial costs associated with using on-phone interpreters is approximately four times that of conducting a similar interview in English. Not only are there the additional costs of the interpreter, but the interviews took about twice as long to complete. The cost of these interviews, however, needs to be compared with the costs and logistical challenges of hiring, training, and supervising interviewers in each of these languages, some of which occur quite infrequently.

Researchers need to develop survey designs that better address the increasingly complex linguistic mix of the US population. Real-time interpretation is one approach for expanding the reach of telephone surveys beyond those who speak only English or Spanish, ensuring that the opinions, needs, and behaviors of those who speak other languages are appropriately accounted for in survey statistics.

5. References

American Association for Public Opinion Research. (2004). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys (3rd ed). Ann Arbor, Mich.: AAPOR.

Eyton, J., and Neuwirth, G. (1984). Cross-cultural Validity: Ethnocentrism in Health Studies with Special Reference to the Vietnamese. Social Science and Medicine, 18, 447–453.

Hilton, A., and Skrutkowski, M. (2002). Translating Instruments into Other Languages: Development and Testing Processes. Cancer Nursing, 25, 1–7.

Mokdad, A.H., Stroup, D., and Giles, H.W. (2003). Public Health Surveillance for Behavioral Risk Factors in a Changing Environment: Recommendations from the Behavioral Risk Factor Surveillance Team. Morbidity and Mortality Weekly Report, 52 (No. RR-9),1–12.

Murray, M.C., Battaglia, M., and Cardoni, J. (2004). Enhancing Data Collection from "Other Language" Households. In Eighth Conference on Health Survey Research Methods, S.B. Cohen and J.M. Lepkowski (eds). Hyattsville, Md.: National Center for Health Statistics, 109–114..

Table 1. Language of interviews

Language	n	Language	n
Vietnamese	37	Turkish	2
Mandarin	31	Ukranian	2
Cantonese	26	Arabic	1
Korean	24	Bosnian	1
Russian	20	French	1
Armenian	10	Hindi	1
Japanese	9	Hirudhi	1
Farsi	5	Hmong	1
Punjabi	5	Indonesian	1
Tagalog	5	Lao	1
Cambodian	3	Portuguese	1
Amharic	2	Thai	1
Bengali	2	(Total)	(195)
Persian	2		

Table 2. Comparison of demographic characteristics, language follow-up and 2005 CA BRFSS

Characteristics	BRFSS % (n)	95% CI	Language Pilot % (n)	95% CI
Sex				
Male	40.9	39.3, 42.4	46.9	39.5, 54.5
Female	59.1	57.6, 60.7	53.1	45.5, 60.5
[n]	[6,134]		[195]	
Age				
18 – 34	28.7	27.3, 30.2	6.9	3.9, 11.8
34 – 54	32.7	31.3, 34.2	25.4	19.3, 32.6
55 – 64	23.3	22.1, 24.6	30.1	23.3, 37.9
65+	15.2	14.3, 16.2	37.7	30.9, 45.0
[n]	[6,131]		[195]	
Race/ethnicity				
Hispanic	34.9	33.4, 36.5	---[1]	---
White, non-Hispanic	53.1	51.6, 54.6	20.4	15.2, 26.8
Black, non-Hispanic	4.4	3.8, 5.1	---[1]	---
Asian	6.1	5.4, 6.9	72.3	65.3, 78.3
Other race/ethnicity	1.5	1.1, 1.9	7.3	4.5, 11.8
[n]	[5,982]		[195]	
Education				
Less than high school	17.9	16.6, 19.3	32.9	26.4, 40.2
High school diploma / GED	22.5	21.2, 23.9	30.1	23.4, 37.8
Some college or more	59.6	58.0, 61.1	37.0	30.1, 44.4
[n]	[6,096]		[195]	
Income				
< $25,000	30.9	29.3, 32.4	62.2	53.2, 70.4
$25,000 - $49,999	24.4	23.1, 25.8	20.6	14.3, 28.8
$50,000 - $74,999	15.2	14.1, 16.3	6.5	2.9, 13.9
$75,999+	29.6	28.2, 31.0	10.7	6.0, 18.2
[n]	[5,628]		[139]	
Number of adults in household				
One	14.2	13.4, 15.0	10.4	7.6, 14.1
Two	51.9	50.3, 53.4	48.3	40.9, 55.9
Three	33.9	32.3, 35.6	41.2	33.5, 49.4
[n]	[6,134]		[195]	
Number of children in household				
None	53.6	52.0, 55.1	55.0	47.4, 62.4
One or more	46.4	44.9, 48.0	45.0	37.6, 52.6
[n]	[6,134]		[195]	

[1] Included in "other race/ethnicity" category: Hispanics (n = 1), black, non-Hispanic (n = 2).

Table 3. Question-level interpretation quality assessment

Measures	Error interpreting question	Error relaying response	Concepts were difficult to translate accurately	Question repeated one or more times	Side conversations between interpreter and respondent
			Proportion of responses per question with a problem		
Mean across all questions	3.6	1.4	4.0	10.7	4.4
Poition in questionnaire:					
Q1 – Q25	3.0	1.6	4.7	11.7	5.5
Q26 – Q50	3.6	1.8	4.2	11.1	4.4
Q51 – Q76	4.2	0.8	3.1	9.3	3.4
(p-value)	(.291)	(.035)	(.150)	(.148)	(.432)
Primary or follow-up question:					
Primary	3.2	1.5	5.4	11.7	5.4
Follow-up	4.1	1.2	2.4	9.6	3.2
(p-value)	(.155)	(.336)	(<.001)	(.187)	(.008)
Type of question:					
Yes/no format	3.4	0.8	3.8	7.4	2.6
Categorical	5.0	2.9	5.4	15.9	6.2
Numeric	3.6	1.6	3.8	13.6	5.7
Likert	3.4	1.6	3.9	12.2	5.9
(p-value)	(.524)	(.004)	(.585)	(<.001)	(<.001)
Questionnaire section:					
Health Status	4.3	2.7	4.2	16.5	9.6
Healthy days	2.5	2.2	5.7	24.0^1	12.8^1
Health care status	2.7	1.6	5.6	16.0	7.6
Exercise	3.2	2.1	6.8	17.4	7.4
Diabetes	0.5	1.1	3.1	4.8	2.2
Oral health	2.5	2.3	2.8	14.2	8.0
Cardiovascular disease	6.4	0.9	7.9	5.9	3.6
Asthma	1.9	2.5	1.3	6.0	0.8
Immunization	3.0	0.9	4.6	6.5	1.9
Tobacco use	2.6	0.8	2.9	7.1	1.4
Alcohol consumption	7.4^1	3.0^3	3.6	14.4	5.1
Demographics	2.9	1.6	4.6	10.2	4.1
Disability	2.0	0.3	2.9	10.2	4.7
Falls/injuries	1.0	0.4	1.6	10.7	4.7
Seatbelt use	3.6	0.9	4.7	7.3	1.5
Mammogram	4.0	0.5	2.3	11.7	4.5
Pap smear	5.3	0.3	2.6	5.2	1.3
Prostate test	4.2	0.9	2.1	8.4	2.4
Colon cancer screening	2.8	0.0^3	2.2	4.1	2.3
HIV/AIDS	5.1	1.4	2.7	13.1	4.1
Emotional support	7.9^2	3.5^4	9.9^5	22.5^6	9.0

(n = 75)   [1] p<.001, [2] p=.019, [3] p=.009, [4] p=.042, [5] p=.003, [6] p=.014

Table 4. Interview-level interpretation quality assessment

Measures	n	Error interpreting question	Error relaying response	Concepts were difficult to translate accurately	Question repeated one or more times	Side conversations between interpreter and respondent
				Proportion of questions per interview with a problem		
Mean across all interviews	195	3.1	1.4	5.3	11.2	5.2
Sex						
Male	86	2.4	1.1	6.0	8.6	4.4
Female	109	3.6	1.7	4.8	13.3	5.9
(p-value)		(.188)	(.396)	(.388)	(.044)	(.274)
Age (years)						
18 – 49	60	4.0	1.5	5.1	10.8	2.3
50 – 69	75	2.2	1.6	4.8	9.5	4.7
70 or older	60	3.2	1.2	6.1	13.9	8.8
(p-value)		(.320)	(.850)	(.731)	(.276)	(<.001)
Education						
Less than high school	67	2.9	2.0	4.2	9.0	4.6
High school diploma/GED	54	3.9	1.1	6.4	12.5	6.5
Some college or more	74	2.6	1.1	5.5	12.4	4.9
(p-value)		(.524)	(.369)	(.440)	(.371)	(.511)
Annual family income						
Less than $15,000	59	2.7	2.1	4.9	10.9	6.0
$15,000-$24,999	33	1.3	1.5	3.0	14.0	5.1
$25,000 or more	47	3.9	2.6	7.5	10.7	5.0
(p-value)		(.212)	(.458)	(.154)	(.621)	(.854)
Language						
Vietnamese	37	0.4	0.1	2.2	4.1	1.0
Mandarin	31	1.3	0.2	3.6	12.7	2.3
Cantonese	26	1.6	0.6	0.4	6.3	1.0
Korean	24	1.7	0.1	4.5	14.6	0.6
Russian	20	0.8	0.3	5.4	18.1	18.9
Other	57	8.1	4.3	10.8	13.5	8.7
(p-value)		(<.001)	(<.001)	(<.001)	(<.001)	(<.001)

Significance based on F-test of means.