Culture and Survey Question Answering: A Behavior Coding Approach

Young Ik Cho, Anne Fuller, Thom File, Allyson L. Holbrook, and Timothy P. Johnson
Survey Research Laboratory, University of Illinois at Chicago

Abstract

Creating survey questions that are well understood uniformly across cultures is a challenge of growing importance for survey researchers conducting surveys of increasingly heterogeneous populations. A number of studies have demonstrated that respondent comprehension of survey questions varies across cultures (e.g., Warnecke et al., 1997). One approach to doing so has been to use behavior coding, a methodology in which the overt verbal behaviors of interviewers and respondents that might indicate problems with a survey question are coded. Although more commonly used in pretesting survey instruments to identify problem questions for revision, this methodology also has been used successfully to assess difficulties that members of different cultural groups may have with understanding and answering survey questions (e.g., Zahnd et al., 2005; Johnson et al., 2006). This evidence suggests that members of minority groups may have more difficulty overall with survey questions, but that cultural differences may not be equally large for all types of survey questions. Our research builds upon this work to examine an extended set of respondent behaviors that may indicate difficulties with various aspects of understanding and answering survey questions among members of three cultural groups (non-Latina White, Latina, and African American). To do so, we use behavior coding data from a face-to-face survey in which 93 women over the age of 40 from these three groups were asked a series of 40 questions about their experiences, behaviors, and beliefs related to cancer and cancer screening. Our goal was to test whether there were cultural differences in the processes involved in understanding and answering these questions and the extent to which cultural variation in these processes could be predicted by the type of survey question involved. Hierarchical linear modeling was used to conduct these analyses, which controlled for other potential sources of variation in response behavior.

Keywords: question comprehension, question mapping, behavior coding, multilevel analysis

1. Introduction

Reducing health disparities is a priority of the National Institutes of Health (Haynes & Smedley, 1999; Smedley, Stith, & Nelson, 2003). Much of the information used to assess these disparities currently is derived from epidemiological surveys that are known to be subject to a variety of sources of error (Groves, 1989) such as differential measurement across varying cultural groups of respondents (Stewart & Nápoles-Springer, 2003). Previous work employing a variety of techniques provides evidence consistent with the differential measurement hypothesis, including cognitive interviewing (Johnson et al., 1997; Warnecke et al., 1997) and item response theory (Morales, Reise, & Hayes, 2000). More recently, behavior coding has also been successfully utilized to investigate cultural variability in respondent answers to health survey questions. Behavior coding is a methodology by which the overt verbal behaviors of interviewers and respondents that might indicate problems with survey questions are coded and systematically analyzed (Fowler & Cannell, 1996; Sykes & Morton-Williams, 1987). An important advantage of this approach is that it introduces objective information beyond respondent answers. Analyses of question behavior codes have revealed general cross-group differences in the comprehension of health questions (Holbrook, Cho, & Johnson, in press; Johnson et al., 2006). These analyses have demonstrated that White respondents exhibit fewer comprehension difficulties when answering health questions, compared to members of minority groups, including African-American, Mexican-American, and Puerto Rican respondents. The reasons for these differences remain unclear, although we can speculate that White respondents may be more likely to share cultural background with the researchers who designed the questionnaires, relative to minority respondents, thereby giving them an advantage in being able to successfully process the questions. These findings have important implications for the design and analysis of health surveys in the United States, as they suggest the possibility that cross-group comparisons may be biased by nonequivalent measurement tools. Given the potential importance of this finding, verifying that it is consistent and replicable across studies is the first objective of this paper.

Previous findings from studies employing behavior coding analyses also suggest that several question design features may be associated with cognitive difficulties when answering questions. After analyzing behavior codings from the National Survey of Recent College Graduates, Cahalan, Mitchell, Gray, and Chen (1994) concluded that certain types of question formats were more likely to be associated with problematic codes, including long questions, those asking about sensitive behaviors, introductory questions to items in a series, questions asking for detailed information, and questions asking about information that might be difficult to recall. Using more quantitative analytic methods, Holbrook et al. (in press) identified more difficult question reading levels, more abstract questions, numeric response formats, and the use of qualified judgments as question characteristics associated with comprehension difficulties. In addition, longer questions, those with more difficult reading levels, numeric response formats, and use of qualified judgments were found to be question

characteristics related to mapping difficulties. These findings provide valuable quantitative evidence that contributes to current knowledge regarding the specific qualities of survey questions that can be expected to systematically produce greater difficulty for respondents and, hence, greater likelihood of measurement error. These findings, however, also warrant replication given the relatively small amount of empirical literature currently available and the importance of these findings. Consequently, the second objective of this paper is to investigate the degree to which earlier findings can be confirmed in a replication study. In addition, we expand the set of question design characteristics to be examined to include requests for subjective vs. objective judgments, requesting information about sensitive topics, the use of showcards, the use of qualified definitions, and the effects of the serial or repetitive use of a question response format.

2. Methods

2.1 Study Sample

A total of 119 face-to-face interviews were conducted in Chicago for a pilot study on respondent experiences, behaviors, and beliefs related to cancer and cancer screening. Of these 119, consent was obtained from 95 respondents to audiotape the interview. Respondents were African-American *(n*=33), Latina *(n*=31), and non-Latina White *(n*=31) women ranging in age from 41–76. They were recruited via advertisements in local newspapers, on www.craigslist.com, and through flyers posted in Chicago. Respondents also were urged to tell their friends and family about the study. The interviews lasted approximately 50 minutes, and respondents were paid $50 for their participation. One African-American female interviewer conducted 117 interviews. A White female served as the interviewer for 2 additional cases. All interviews were conducted in the Chicago offices of the University of Illinois at Chicago Survey Research Laboratory.

2.2 Coding Question Characteristics

The survey instrument contained 84 questions. Of these, we behavior coded responses to the 40 questions that were asked of all respondents. These questions were classified along nine dimensions: question length, reading difficulty level, response format, number of scale points, use of a showcard, use of objective or subjective judgments, use of qualified definitions, question sensitivity, and abstraction level. The specific wording of each survey question and coding for each dimension are available from the authors by request. Question length was measured by total number of words. The school grade reading level of each question was measured using Flesch-Kincaid scores (Flesch, 1979). Three response formats were included: those for which the respondent could answer “yes” or “no” or “true” or “false”; those employing Likert-type response scales (including both unipolar and bipolar scale verbal labels); and those for which the respondent responded with a number. The questions also were classified as to whether or not they used showcards. Showcards provide a visual aid for respondents by listing response options on paper. Eleven of the 40 questions used showcards.

Two of the authors independently coded the questions according to use of objective or subjective judgments, question sensitivity, use of qualified definitions and level of abstraction. Objective judgments were defined as those involving information about a respondent’s behavior (e.g., “Have you ever had a mammogram?”). Subjective judgments were defined as those involving information about a respondent’s beliefs, attitudes, or other similar type of subjective judgment (e.g., “In general, I trust my doctor to give me the best possible health care. Would you say this is always true, mostly true, half the time true, sometimes true or never true?”). Qualified definitions were defined as those involving a specified time frame (e.g., during the past year) or excluding items from a category (e.g. “servings of vegetables, not counting salads or potatoes”). Initial agreement between the two coders on these three dimensions was very high, and differences were discussed and reconciled.

Questions also were classified as either sensitive or not sensitive. Questions that were not sensitive were those that would not cause discomfort for the average respondent (e.g., “In general, would you say your health is excellent, very good, good, fair, or poor?”). Questions that were classified as sensitive were those that might cause discomfort for some respondents (e.g., “Have you ever smoked marijuana?”). Three levels of abstraction were used to classify the questions. The levels were “least abstract,” “somewhat abstract,” and “most abstract.” Items were defined as “least abstract” if the major concept introduced in the question was grounded in a physical reality (e.g., “pap smear” and “mammogram”). Items were defined as “most abstract” if the major concept introduced by the question was not grounded in physical reality (e.g., “control over future health” or “stress”). The remaining items were classified as “somewhat abstract” and introduced moderately abstract concepts (e.g., “eating salad” or being a “regular smoker”). The coders agreed on a majority of the items and differences were discussed and reconciled.

2.3 Behavior Coding

The same two authors behavior-coded respondent reactions to each of the 40 survey questions. As many as three codes could be assigned to each question response. The behavioral data coded for this analysis included respondent reactions when each question was asked and when the answer was recorded (many questions were then followed up with one or more structured probes, but behavioral responses to these probes were not coded or analyzed). Overall, 7,980 respondent answers were coded.

On two occasions, five taped interviews were coded by both of the research assistants, for a total of 10 tapes and 840 total responses. The inter-rater agreement in both instances was high. For the first five interviews coded, the level of agreement was 92.6%. The differences between the two coders were discussed and reconciled. For the next five interviews coded, the percent agreement was 96.6%. The remaining 85 interviews were subsequently coded independently by the coders.

Behavior codes were subsequently employed to construct summary indicators of the presence or absence of comprehension and mapping difficulties. Comprehension problems may occur when question language or concepts are poorly fitted to the way a respondent thinks about the subject in question. Behavior codes that suggest that respondents may not have understood the basic objective of the question might be expected to represent comprehension problems. Mapping difficulties are dependent on respondent ability to simultaneously keep a question in mind while selecting an appropriate response to it. Hence, the question's response format may influence mapping problems. Based on previous work reported by Holbrook et al. (in press), summary indicators of comprehension vs. mapping difficulties were constructed by coding question answers as reflecting one of these constructs if any one of five specific behavior codes representing each of these potential sources of respondent difficulty were associated with that response. The specific behavior codes associated with each source of respondent difficulty are reproduced elsewhere (see Holbrook et al., in press).

2.4. Analysis

We employed hierarchical linear modeling (HLM) to estimate two-level models that examined the variance attributable to both individual-level and question-level characteristics (Raudenbush, Bryk, Cheong, & Congdon, 2004). The individual-level characteristics examined were race/ethnicity (African American vs. Latina vs. white), age, and education. The nine question-level characteristics described above were also examined.

3. Results

3.1 Sample Description

The distributions of the two summary cognitive difficulties indicators, and question characteristics, among the responses included in this analysis (*n*=3,720) are presented in the upper panel of Table 1. Overall, responses to 5% of all survey questions were coded as having comprehension difficulties associated with them. Mapping difficulties were associated with 3% of all responses obtained. The proportions of each survey question that were classified as having comprehension difficulties ranged from 0–17%. For mapping difficulties, the proportion of questions so identified ranged from 0–19%**.** These values suggest that most of the items examined fell below the general cut-off rule that suggests that items eliciting 15% or more problematic behavior codes should be considered poor-performing items (Zukerberg, Von Thurn, & Moore, 1995).

The mean length of the survey questions responded to was 16.5, with a standard deviation of 7.6, suggesting considerable heterogeneity in question length. The mean reading level of the questions answered was grade 6.3, with a standard deviation of 2.4, again suggesting the sample had variability in question-reading difficulty. Overall, the questions were primarily subjective (75%) and employed a yes-no response format (63%). Few employed showcards (28%), were deemed sensitive (15%) or included a qualified definition (10%). Few questions requested numeric responses (5%). Twenty-eight percent were coded as falling in the "most abstract" category, 33% were coded as being "somewhat abstract," and 40% were coded into the "least abstract" category. On average, each question was immediately preceded by 4.8 questions that employed the same response format. The wide standard deviation (4.9) indicates there is again considerable variability in the numbers of questions using the same response format preceding each item.

3.2 Hierarchical Models

The lower panel of Table 1 indicates that, by design, the sample (*n*=93) was composed of nearly identical proportions of African American (34%), Latina (33%), and White (32%) respondents. The mean age of the sample was 51.9 years (standard deviation=7.9). Respondent age ranged from 41–77 years old. The sample was fairly evenly distributed in terms of educational attainment. Thirty-seven percent reported a high school or less education. A third indicated having some college education, and 30% were college graduates.

Table 1. Descriptive Statistics

Variable	Mean	SD	Minimum	Maximum
		Question Level (*N*=3,720)		
Comprehension Difficulty	0.05	0.22	0	1
Mapping Difficulty	0.03	0.16	0	1
Number of Words	16.48	7.59	5.0	38
Reading Level	6.27	2.44	0.8	11.7
Qualified Definition	0.1	0.3	0	1
Show Card	0.28	0.45	0	1
Subjective Judgment	0.75	0.43	0	1
Sensitive Question	0.15	0.36	0	1
Response Format				
Yes/no	0.63	0.48	0	1
Likert-scale	0.33	0.47	0	1
Numerical	0.05	0.22	0	1
Abstraction Level				
Least abstract	0.40	0.49	0	1
Somewhat abstract	0.33	0.47	0	1
Most abstract	0.28	0.45	0	1
Consecutive Question Format Repetition	4.78	4.92	0	16
		Person Level (*N*=93)		
Age	51.85	7.86	41.0	77
Ethnicity/Race				
Latina	0.33	0.47	0	1
African American	0.34	0.48	0	1
White	0.32	0.47	0	1
Education				
High school graduate or less	0.37	0.48	0	1
Some college	0.33	0.47	0	1
College graduate	0.30	0.46	0	1

The first equation in Table 2 presents the results of an HLM model predicting comprehension difficulties for the sample of 3,686 survey responses. Latina respondents were found to be more likely than Whites to express comprehension difficulties, net of the other variables included in this analysis. Overall, 3.8% of all answers given by White respondents were coded as having comprehension difficulties. In contrast, 6.3% of all answers given by Latina respondents featured comprehension problems, and 4.3% of all responses provided by African Americans were coded as having comprehension problems. The adjusted mean difference in comprehension problems between African-American and White respondents was not significant. Equation 1 in Table 2 also revealed a positive association between respondent age and comprehension difficulties.

Four question characteristics also were found to be associated with comprehension difficulties. Increasing question length and increasing reading level were both associated with more comprehension problems. In addition, questions that included qualified definitions were less likely to lead to difficulties in comprehension. Compared to questions that required a numeric response, questions using yes-no and Likert-type response formats were also less likely to produce comprehension problems. The adjusted relationships between each question characteristic and comprehension problems are depicted in Figure 1.

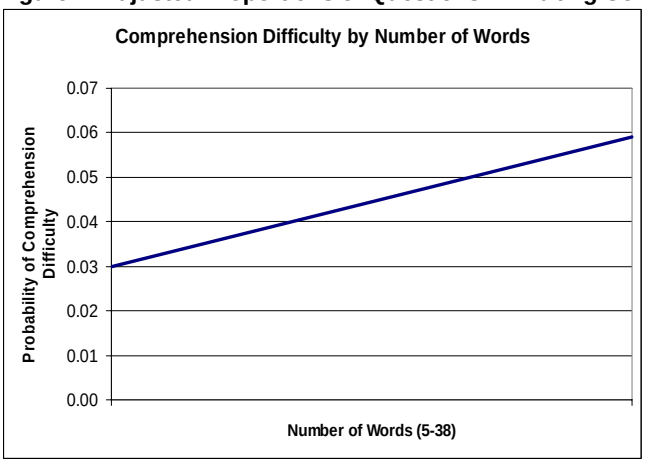
Equation 2 in Table 2 examined the effects of question and respondent characteristics on the likelihood of expressing mapping difficulties. Also, age was positively correlated with this outcome measure. In addition, the responses of African-American respondents were more likely to indicate mapping difficulties than those of White respondents. Of all responses by African Americans, 2.3% were identified as having mapping difficulties, compared to 1.1% of the answers provided by White respondents and 2.0% of the answers provided by Latina respondents. The difference between Latina and White respondents was borderline (i.e., $p<.10$) significant in this analysis. This second equation (Table 2) also revealed independent associations between mapping difficulties and four question characteristics. Responses to questions with higher levels of reading difficulty were more likely to be associated with mapping problems in these data, as were responses to questions requiring a subjective judgment. Answers to questions that requested yes-no or Likert-type response formats were again more likely to produce fewer problems, relative to responses to questions that required numeric answers. Additionally, there were fewer expressions of mapping difficulties to questions that were positioned within the instrument such that greater numbers of prior questions employed the same response format. Figure 2 presents the adjusted relationships between each of these question characteristics and mapping problems.

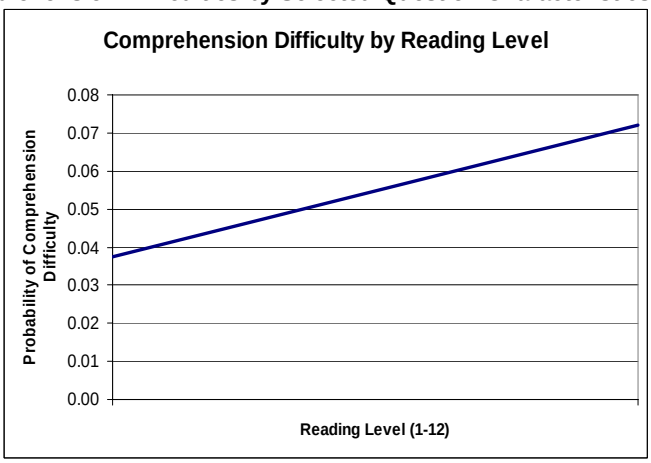
Table 2. Hierarchical Model of Effects of Person- and Question Level Variables on Comprehension and Mapping Difficulty

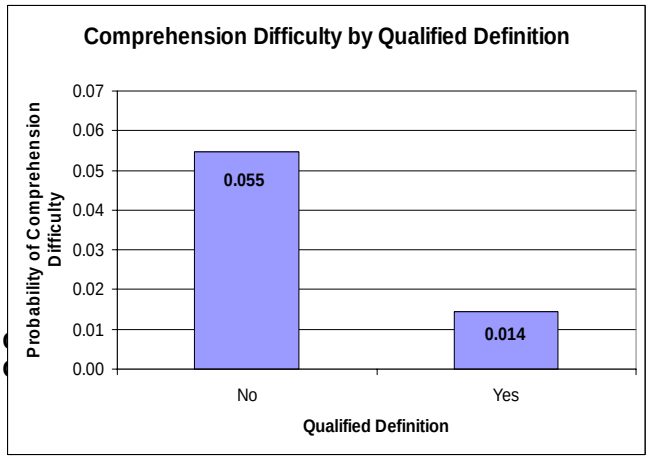
	EQUATION 1 Comprehension Difficulty (1=yes)		EQUATION 2 Mapping Difficulty (1=yes)	
Level 1 df	3,686		3,686	
Level 2 df	87		87	
	Coefficient	**(S.E.)**	**Coefficient**	**(S.E.)**
Intercept	-2.99***	(0.11)	-4.05***	(0.16)
Question Characteristics				
Number of words	0.03*	(0.02)	-0.03	(0.02)
Reading level	0.11*	(0.05)	0.21**	(0.06)
Qualified definition	-1.38*	(0.68)	-0.62	(1.25)
Show card	0.69+	(0.39)	-0.47	(0.45)
Subjective judgment	0.57	(0.40)	2.46***	(0.46)
Sensitive question	0.04	(0.39)	0.0004	(0.35)
Response format (Ref=Numerical)				
Yes/no	-1.77*	(0.73)	-2.18*	(1.06)
Likert	-2.96**	(0.95)	-3.27*	(1.31)
Abstraction level (Ref=Most abstract)				
Least abstract	-0.47	(0.32)	-0.20	(0.33)
Somewhat abstract	-0.18	(0.23)	0.38	(0.31)
Consecutive question format repetition	-0.03	(0.02)	-0.19***	(0.04)
Person Characteristics				
Age	0.02**	(0.01)	0.05**	(0.02)
Ethnicity/Race (Ref=White)				
Latina	0.51*	(0.20)	0.65+	(0.38)
African American	0.14	(0.25)	0.77**	(0.28)
Education (Ref=College Graduate)				
High school graduate or less	0.38	(0.23)	0.03	(0.29)
Some college	0.25	(0.21)	0.22	(0.39)

***p<.001; **p<.01; *p<.05; +p<.10.

Figure 1. Adjusted Proportions of Questions Exhibiting Comprehension Difficulties by Selected Question Characteristics







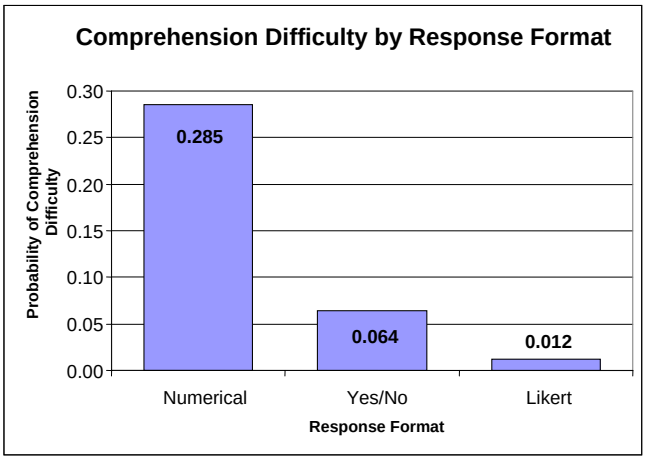
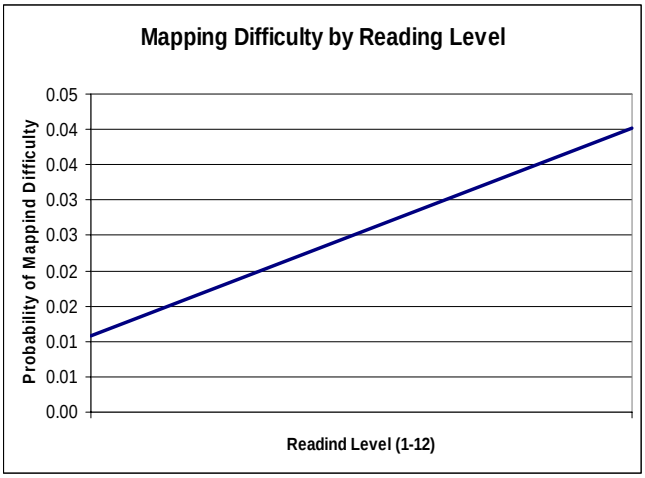
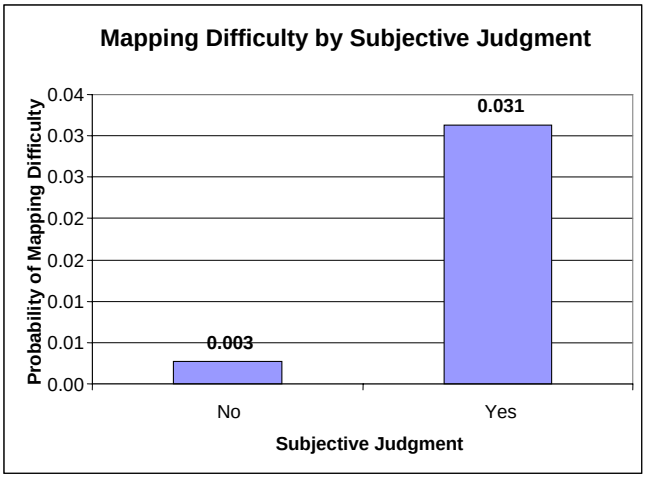
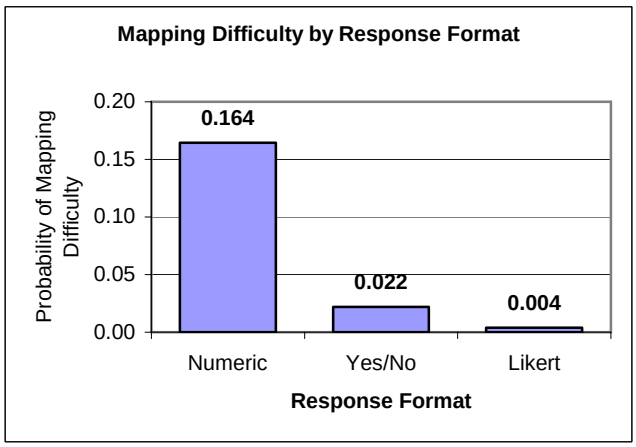
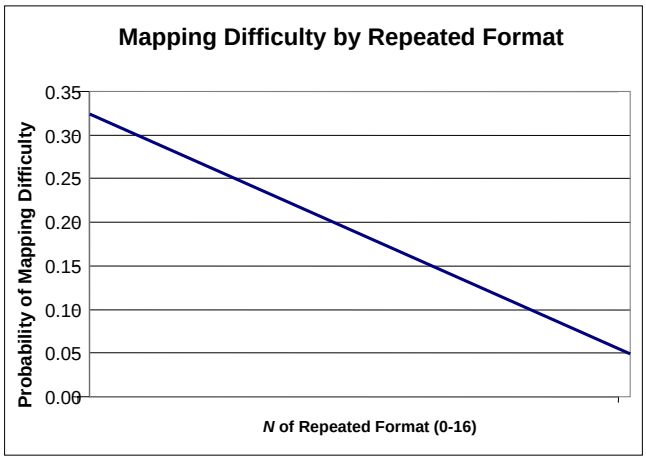


Figure 2. Adjusted Proportions of Questions Exhibiting Mapping Difficulties by Selected Question Characteristics









4. Discussion

4.1 Effects of Culture on Processing Difficulties

The first objective of this investigation was to attempt to replicate previous findings of a relationship between respondent culture and cognitive processing difficulties when answering survey questions. Partially consistent with previous work reported by Holbrook et al. (in press), one of two minority populations examined in this study (Latinas) was found to express more comprehension difficulty when answering a set of 40 survey questions concerned with cancer screening beliefs and behaviors. Unlike this previous research, no differences in comprehension problems were found between African-American and White respondents, although we note that (a) the direction of the association was nonetheless in the same direction (i.e., with larger numbers of problems expressed by African Americans relative to Whites), and (b) the differences in comprehension problems between Latina and White respondents were greater in both studies, relative to the African American-White differences. It may be that the smaller sample size employed in the current study *(n*=93 total interviews vs. *n*=345 total interviews in that previous research) resulted in insufficient power to detect the smaller cross-group differences between African-American and White respondents.

Unlike the previous research by Holbrook et al. (in press), which did not detect cultural variability in mapping difficulties, cross-group differences were identified in the current study. Specifically, African Americans were more likely than White respondents to express mapping problems, and the direction of the association was similar, albeit only borderline significant, for Latina respondents. To our knowledge, this is the first study to identify racial/ethnic variability in question mapping tasks.

4.2 Effects of Question Characteristics on Processing Difficulties

The second objective of this study was to replicate previous findings regarding associations between a set of common question characteristics and respondent processing difficulties. Consistent with the Holbrook et al. study (in press), question-reading level was found to be positively associated with comprehension problems. Also consistent was the finding that numeric response formats were more problematic in regards to respondent comprehension than were yes-no and Likert-type response formats. Other findings, however, did not replicate. Unlike the Holbrook et al. (in press) results, question length (i.e., total number of words) was positively associated with comprehension difficulty. Because longer questions are likely to require more working memory, it is perhaps not surprising to find them associated with greater problems of comprehension. Given the older age of our sample (mean=51.9) relative to Holbrook's study (mean=32.1), it may be that the effects of question length on comprehension problems is more pronounced among older respondents due to their declining cognitive resources (Schwarz, Park, Knäuper, & Sudman, 1999). Also, we failed to replicate previous findings (Holbrook et al., in press) between level of question abstraction and comprehension problems, although we note the direction of the relationship was consistent, with increasing levels of abstraction associated with increasing difficulties in comprehension. The current study also found the presence of qualified definitions to be negatively related to comprehension difficulties. The use of showcards, requesting subjective vs. objective judgments, requesting sensitive vs. nonsensitive judgments, and response format repetition were not found to be associated with comprehension problems.

The associations between question features and mapping difficulties were also only partially replicated. In both studies, question-reading level was positively associated with mapping problems, and numeric response formats also led to greater problems, relative to yes-no and Likert-like formats. Question length and the use of qualified definitions, though, were not associated with mapping difficulties in the present study. Each had been previously reported by Holbrook et al. (in press) as being related to mapping problems. One question format not previously examined, requesting subjective (vs. objective) assessments, was found to be associated with mapping difficulties in the current study. The use of showcards and requests for sensitive vs. nonsensitive judgments were not related to mapping difficulties.

4.3 Study Limitations

We are aware of several limitations of this study. The study is based on a relatively small number of respondents (*n*=93). Thus, as already suggested, there may be insufficient power to detect some group differences. In addition, the study is restricted to females age 40 and older from three cultural groups, which limits its generalizability. Also, one interviewer was responsible for completing 91 of the 93 interviews included in these analyses, further making it difficult to generalize beyond respondent interactions with this individual. In regards to behavior coding, it is important to remember that this methodology was initially developed to evaluate interviewer behavior, not investigate respondent cognitions. It is also unclear whether the inherent assumption that respondents from varying cultural backgrounds overtly express the behaviors being coded in a similar manner and to a similar degree is appropriate. Ultimately, however, we believe the use of behavior coding is an important strength of this study, as it relies on objective assessments of respondent behaviors to evaluate difficulties in the cognitive processing of survey questions. In addition, the use of HLM modeling is another strength of this study, as it facilitates appropriate modeling of the effects of both respondent- and question-level characteristics on the outcomes of interest.

5. Conclusions

These findings provide evidence supportive of earlier research that suggests the presence of systematic variations in health survey question comprehension across several cultural groups. As such, it contributes to mounting evidence that standardized survey interviewing protocols may be insufficient to insure measurement equivalence across multiple cultural groups when conducting epidemiological and other health-related research. Of course, given the limitations cited above, this study is far from definitive. It does nonetheless contribute to mounting evidence of culture-based variability in the cognitive processing of survey questions that requires further evaluation and resolution. These findings also confirm earlier research that has documented variability in survey question processing that appears to be related to elements of the questions themselves. Although these findings are not completely consistent with the earlier research, there is enough consistency, given this study's design, to conclude that additional research is necessary to further elaborate the implications of this research for best questionnaire design practices.

Acknowledgements

Support for this research was provided by grant P50-CA106743 from the National Cancer Institute

References

Flesch, R. F. (1979). *How to write plain English.* New York: Harper and Row.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey instruments. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research*, (pp.15–36). San Francisco: Jossey-Bass.

Cahalan, M., Mitchell, S., Gray, L., & Chen, S. (1994). *Recorded interview behavior coding study of national survey of recent college graduates.* Paper presented at the annual meeting of the American Statistical Association.

Groves, R. M. (1989). *Survey errors and survey costs.* New York: John Wiley & Sons.

Haynes, M. A., & Smedley, B. D. (1999). *The unequal burden of cancer: An assessment of NIH research and programs for ethnic minorities and the medically underserved.* Committee on Cancer Research Among Minorities and the Medically Underserved. Washington, D.C.: Institute on Medicine, National Academy Press.

Holbrook, A. L., Cho, Y. I., & Johnson, T. P. (in press). The impact of question and respondent characteristics on comprehension and mapping difficulties.

Johnson T. P., Cho Y. I., Holbrook A., O'Rourke D., Warnecke R. B., & Chávez, N. (2006). Cultural variability in the effects of question design features on respondent comprehension of health surveys. *Annals of Epidemiology,* 15, 661-668.

Johnson T. P., O'Rourke, D., Chávez, N., Sudman, S., Warnecke, R., Lacey, L., et al. (1997). Social cognition and responses to survey questions among culturally diverse populations. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and process quality* (pp. 87–113). New York: John Wiley & Sons.

Morales, L. S., Reise, S. P., & Hays, R. D. (2000). Evaluating the equivalence of health care ratings by Whites and Hispanics. *Medical Care*, 38, 517–527.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F, & Congdon, R. T. (2004*). HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (1999). *Cognition, aging, and self-reports*. Philadelphia: Psychology Press.

Smedley, B. D., Stith, A. Y., & Nelson, A. R. (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, D.C.: The National Academies Press.

Stewart, A. L., & Nápoles-Springer, A. M. (2003). Advancing health disparities research: Can we afford to ignore measurement issues? *Medical Care, 41*, 1207–1220.

Sykes, W., & Morton-Williams, J. (1987). Evaluating survey questions. *Journal of Official Statistics, 3*, 191–207.

Warnecke, R. B., Johnson, T. P., Chávez, N., Sudman, S., O'Rourke, D. P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7*, 334–342.

Zahnd, E., Tam, T., Lordi, N., Willis, G., Edwards, W. S., Fry, S. et al. (2005, July). *Cross-cultural behavior coding: Using the 2003 California Health Interview Survey to assess cultural/language data quality.* Paper presented at the first European Association for Survey Research Conference, Barcelona, Spain.

Zukerberg, A. L., Von Thurn, D. R., & Moore, J. C. (1995). *Practical considerations in sample size selection for behavior coding pretests.* Pp. 1116–1121 in the Proceedings of the Annual Meeting of the American Statistical Association. Alexandria, VA: American Statistical Association.