ARTICLES

Do Interviewer Errors Help Explain the Impact of Question Characteristics on Respondent Difficulties?

Allyson L. Holbrook, Timothy P. Johnson, Young Ik Cho, Sharon Shavitt, Noel Chavez, Saul Weiner

Keywords: survey practice

<https://doi.org/10.29115/SP-2016-0009>

Survey Practice

Vol. 9, Issue 2, 2016

Respondents, interviewers, and questions are all sources of potential error in responses to survey questions. Using behavior coding to measure respondent problems and interviewer errors, we assess whether question characteristics influences these behaviors. We also assess whether interviewer errors may help to explain the impact of question characteristics on respondent difficulties. Specifically, we assess whether more difficult questions (e.g., longer or more difficult to read) lead to more respondent problems (e.g., answers that do not fit the requirements of the question) at least in part because more difficult questions also lead to interviewer errors (e.g., deviating from the reading the question as written or failing to probe correctly).

Introduction

Researchers have long been concerned about potential error introduced by survey interviewers' behaviors (e.g., van der Zouwen, Dijkstra, and Smit 1991). Survey interviews involve not only the cognitive task of answering survey questions but are also a form of conversation and a social interaction. Behavior coding is one method used to examine the behaviors of respondents and interviewers and the interaction between them (Willis 2005). Behavior coding of respondent and interviewer behaviors has been used in cognitive pretesting to identify questions that need to be revised (Willis 2005), and researchers have also begun to use this approach to study the interaction between interviewers and respondents (e.g., Dykema, Lepkowski, and Blixt 1997) and the cognitive response process respondents go through to answer survey questions (e.g., Holbrook, Cho, and Johnson 2006).

The current research uses behavior coding of respondent and interviewer behaviors to examine how question characteristics affect specific interviewer and respondent behaviors and whether interviewer behaviors account for the effect of question characteristics on respondent behaviors. In our previous research, we found that some respondent behaviors indicate problems with question comprehension (e.g., asking for clarification), whereas others indicate problems with mapping a response onto the response format (e.g., giving an inadequate response) and that different question characteristics are associated with behaviors indicating problems with question comprehension and mapping (Holbrook, Cho, and Johnson 2006). In this paper, we extend this work by examining a larger set of more systematically manipulated question characteristics and by considering the role of interviewer behaviors.

Methods

Respondents and Procedures

Respondents were 603 adults recruited in Chicago. Approximately equal numbers were African Americans, Korean Americans, Mexican Americans and non-Hispanic Whites. Among Korean Americans and Mexican Americans, half of the interviews were in English and half were in Korean and Spanish, respectively. A complete description of the survey procedures can be found in study 2 of Holbrook et al. (2014).

Measures

Behavior coding. Respondents were asked up to 296 different questions in the audio- and video-recorded computer-assisted personal interviewing (CAPI) interview (not all respondents were asked all questions since the questionnaire involved skip patterns). This resulted in approximately 118,000 question-answer sequences across all respondents. For each question-answer sequence, up to three interviewer and three respondent behaviors were coded by coders who listened to the audio files. Twenty percent of question-answer sequences were validated by a second coder (an exact validation rate of 95.7 percent indicated high reliability). Four behavior variables were constructed at the question-answer sequences level. One variable was coded 1 if the respondent demonstrated a behavior that indicated a question comprehension problem (e.g., a request for clarification) and 0 otherwise; a second variable was coded 1 if the respondent demonstrated a behavior that indicated a mapping problem (e.g., a response that did not meet the question objective) and 0 otherwise; the third variable was coded 1 if the interviewer deviated from reading the question exactly as written (either a major or minor deviation) and 0 otherwise; and the fourth variable was coded 1 if the interviewer failed to correctly probe when he or she should have done so and 0 otherwise. A large number of the question-answer sequences showed interviewer reading problems ($n=28,668$; 24.2 percent), respondent comprehension problems ($n=10,737$; 9.1 percent), and respondent mapping problems ($n=14,454$; 12.2 percent). A smaller but still sizeable number showed interviewer probing problems ($n=4,457$; 3.8 percent), perhaps not surprising since an interviewer cannot fail to probe correctly unless probing is necessary which would not occur in every sequence.

Question characteristics. The questionnaire was designed to manipulate several question characteristics including response format (open-ended numeric, agree/disagree, yes/no categorical, unipolar fully-labeled scale, bipolar fully-labeled scale with a midpoint, bipolar fully-labeled scale without a midpoint, semantic differential with only endpoints labeled, semantic differential with endpoints and midpoint labeled, and faces scale); type of judgment (subjective, personal characteristic or behavior, objective factual knowledge); whether the judgment was time-qualified or not (i.e., whether the judgment was restricted to a particular time frame); whether the question asked the respondent to report about him or herself or as a proxy for another family member; whether

the question was designed to be difficult or not (i.e., some questions in the survey were specifically designed to be difficult in a variety of ways); whether or not there was an explicit do not know or no opinion response option included; whether or not the question was preceded by a do not know filter question or not; and whether the question itself was a do not know filter question. In addition, question length was operationalized as the number of sentences and reading difficulty was operationalized as the Flesch-Kincaid grade level. Finally, each question's sensitivity (0=not at all sensitive, 1=somewhat sensitive, 2=very sensitive) and its level of abstraction (0=concrete, 1=moderately abstract, 2=very abstract) were coded by four coders one from each of the four racial and ethnic groups included in the study whose ratings were averaged. Finally, the position of each questionnaire within the CAPI instrument was captured as a measure of the number of previous questions. Questions that were only answered by part of the sample were counted as one half. (The order of sections of the questionnaire was rotated across respondents.)

Respondent characteristics. Our analyses also controlled for respondent characteristics. Interviewers observed respondent gender (coded 0 for females and 1 for males). Age was coded in years and ranged from 18 to 70. Respondents' number of years of education was coded into 7 categories coded to range from 0 to 1 (0=no school; 0.17=6 years or fewer completed; 0.33=7–9 years completed; 0.5=10–12 years completed; 0.67=13–14 years completed; 0.83=15–16 years completed; and 1=more than 16 years completed). Income was coded to into 6 categories coded to range from 0 to 1 (0=less than or equal to \$10,000; 0.20=greater than \$10,000 and less than or equal to \$20,000; 0.4=greater than \$20,000 and less than or equal to \$30,000; 0.6=greater than \$30,000 and less than or equal to \$50,000; 0.8=greater than \$50,000 and less than or equal to \$70,000; and 1=greater than \$70,000). Race/ethnicity and language of interview were coded using five dummy variables (representing non-Hispanic Blacks, Mexican-Americans interviewed in English; Mexican-Americans interviewed in Spanish; Korean-Americans interviewed in English; and Korean-Americans interviewed in Korean with non-Hispanic Whites as the comparison group).

Analysis

Cross-classified multilevel models where dichotomous variables indicating behaviors are the dependent variables that are nested within question and respondent simultaneously were used. This approach corrects for nonindependence at question and respondent levels and can be used to analyze respondent and question characteristics as well as interviewer and respondent behaviors at the question level. Hierarchical linear and nonlinear modeling (HLM) was used to conduct all analyses. These analyses also controlled for question characteristics and respondent demographics.

Results

Results are shown in Table 1 and are presented in the (rough) order in which behaviors occur in the interviewer-respondent interaction. Interviewer reading errors are predicted in model 1. Interviewer reading errors were greater for longer questions and those with higher reading grade levels. They varied somewhat across response formats, but not strongly. Reading errors were also significantly more likely for questions that involved a proxy report, perhaps because these questions tend to be rarer than questions about the respondent him or herself. Interviewer reading errors also increased as the survey progressed, perhaps due to interviewer fatigue. Other question characteristics were unrelated to reading errors.

Respondent comprehension problems are shown in model 2. Comprehension problems were more common for questions with higher reading grade levels, but not for longer questions. Response format had a big impact on respondent comprehension problems with many question response formats (e.g. yes-no, unipolar scales, bipolar scales, and some semantic differential scales) showing fewer respondent comprehension problems than open-ended numeric questions (the baseline category). Respondent comprehension problems were also greater for questions that were designed to be deliberately difficult and time qualified judgments (e.g., those that required respondents to their judgment to a particular time frame). Respondents also had more comprehension problems with “don’t know” filter questions than with other questions – perhaps because it is unusual in everyday conversation to be asked whether you have an opinion about a topic. Proxy questions showed fewer respondent comprehension problems than those asking about the respondent him or herself, although these were all asked late in the interview (the order of the proxy questions was not manipulated) and most paralleled questions asked about the respondent, so this may reflect respondent learning as the interview progresses. Questions that involved a showcard were more likely to result in comprehension problems as well, perhaps because (like answering “don’t know” filter questions) this task also does not parallel everyday conversation. Finally, respondent comprehension difficulties were negatively associated with the number of previous questions – suggesting that respondents learned over the course of the interview. A model predicting respondent comprehension difficulties that includes interviewer reading errors as an independent variable is shown in model 3. Interviewer reading errors were unassociated with respondent comprehension problems.

Respondent mapping problems are predicted in model 4. Higher grade reading levels were associated with more problems, but question length was not associated with mapping problems. As with respondent comprehension problems, almost all the other question types showed fewer mapping problems than open-ended numeric questions. However, agree-disagree questions showed more mapping problems than open-ended numeric questions, providing support for the argument that agree-disagree questions are difficult

Table 1 Multilevel logistic regression models predicting respondent and interviewer behaviors (standard errors shown in parentheses)

Model	Dependent variable						
	I Reading problems	R comprehension problems		R mapping problems		I probing problems	
	1	2	3	4	5	6	7
Question-answer sequence level variables							
Interviewer reading problem			-0.03 (0.03)		-0.001 (0.03)		
Respondent comprehension problem							0.40** (0.05)
Respondent mapping problem							2.74** (0.04)
Question characteristics							
Number of sentences	0.79** (0.07)	0.02 (0.08)	0.02 (0.08)	-0.02 (0.07)	-0.02 (0.07)	-0.05 (0.07)	-0.06 (0.07)
Flesch-Kincaid grade level	0.11** (0.01)	0.04** (0.01)	0.04** (0.01)	0.02* (0.01)	0.03* (0.01)	-0.02* (0.01)	-0.05** (0.01)
Judgment type ^a							
Subjective	-0.39 (0.47)	-0.21 (0.54)	-0.21 (0.54)	0.58 (0.47)	0.58 (0.47)	0.36 (0.52)	0.24 (0.51)
Self	-0.20 (0.52)	-0.73 (0.59)	-0.73 (0.59)	0.35 (0.51)	0.35 (0.51)	-0.20 (0.57)	-0.31 (0.56)
Response format ^b							
Agree-disagree	-0.63* (0.27)	-0.07 (0.31)	-0.07 (0.31)	0.93** (0.25)	0.93** (0.26)	1.19** (0.28)	0.78** (0.27)
Yes-no	-0.21 (0.15)	-0.82** (0.17)	-0.82** (0.17)	-0.88** (0.15)	-0.88** (0.15)	-1.56** (0.17)	1.23** (0.17)
Categorical	-0.26+ (0.15)	-0.34+ (0.18)	-0.34+ (0.18)	-0.20 (0.15)	-0.20 (0.15)	-0.72** (0.18)	-0.62** (0.17)
Unipolar	0.29+ (0.16)	-0.84** (0.19)	-0.84** (0.19)	-0.57** (0.16)	-0.57** (0.16)	-0.56** (0.18)	-0.36* (0.18)
Bipolar with midpoint	0.15 (0.19)	-0.59** (0.22)	-0.59** (0.22)	-0.43* (0.19)	-0.43* (0.19)	-0.48* (0.20)	-0.34+ (0.20)
Bipolar without midpoint	0.49 (0.31)	-0.94* (0.37)	-0.94* (0.37)	0.25 (0.31)	0.25 (0.31)	-0.31 (0.33)	-0.52 (0.33)
Semantic differential with endpoints labeled	-0.01 (0.22)	0.01 (0.26)	0.01 (0.26)	-1.34** (0.23)	-1.34** (0.23)	-0.74** (0.25)	-0.22 (0.25)
Semantic differential with endpoints and midpoint labeled	-0.16 (0.72)	-2.15** (0.84)	-2.16** (0.84)	-2.19** (0.72)	-2.19** (0.72)	1.52+ (0.81)	-0.72 (0.80)
Faces scale	-0.84 (0.57)	-0.10 (0.64)	-0.10 (0.64)	-2.04** (0.56)	-2.04** (0.56)	-1.57** (0.61)	-0.82 (0.60)
Abstraction	-0.05 (0.10)	-0.17 (0.11)	-0.17 (0.11)	-0.23* (0.10)	-0.23* (0.10)	-0.54** (0.11)	-0.50** (0.11)
Deliberately difficult	0.16 (0.37)	2.38** (0.41)	2.38** (0.41)	0.02 (0.36)	0.03 (0.36)	2.42** (0.36)	2.68** (0.35)

	Dependent variable						
	I Reading problems	R comprehension problems		R mapping problems		I probing problems	
Model	1	2	3	4	5	6	7
Time qualified judgement	0.09 (0.09)	0.23* (0.11)	0.23* (0.11)	-0.25** (0.09)	-0.25** (0.09)	-0.20+ (0.11)	-0.11 (0.11)
Explicit DK option provided	-0.32 (0.41)	-0.23 (0.47)	-0.23 (0.47)	-0.17 (0.41)	-0.17 (0.41)	-0.55 (0.46)	-0.51 (0.45)
Preceded by a DK filter question	0.53 (0.58)	-0.75 (0.67)	-0.75 (0.67)	-0.53 (0.58)	-0.53 (0.58)	-0.07 (0.61)	0.06 (0.61)
DK filter question	-0.02 (0.60)	1.87** (0.66)	1.87** (0.67)	1.12+ (0.58)	1.12+ (0.58)	0.90 (0.64)	0.10 (0.63)
Proxy report	0.65** (0.12)	-0.94** (0.15)	-0.94** (0.15)	-0.08 (0.12)	-0.08 (0.12)	0.57** (0.14)	0.72** (0.14)
Sensitivity	-0.19 (0.14)	0.12 (0.16)	0.12 (0.16)	0.38** (0.14)	0.38** (0.14)	0.26+ (0.15)	0.13 (0.15)
Showcard	0.27 (0.42)	1.04* (0.48)	1.04* (0.48)	1.11* (0.45)	1.11* (0.42)	0.67 (0.47)	0.17 (0.44)
Number of previous questions	0.001** (0.0002)	-0.001** (0.0003)	-0.001** (0.0003)	-0.0002** (0.0002)	-0.0002** (0.0002)	-0.001 (0.001)	0.0004 (0.001)
Intercept	-4.93** (0.60)	-2.78** (0.64)	-2.79** (0.64)	-2.70** (0.55)	-2.70** (0.55)	-3.38** (0.64)	4.10** (0.64)
Question-answer sequence level df	107,355	107,355	107,355	107,355	107,355	107,355	107,355
Respondent df	544	544	544	544	544	544	544
Question df	294	294	294	294	294	294	294

All analyses also controlled for respondent gender, race/ethnicity and language of interview, income, education, and age.

^aObjective knowledge questions were the comparison group.

^bQuestions with open-ended numeric response formats were the comparison group.

* $p < 0.05$, ** $p < 0.01$, + $p < 0.10$.

because respondents have difficulty translating their judgments onto the agree-disagree format. Somewhat unexpectedly, more abstract questions and those that were time qualified showed fewer mapping problems, which is inconsistent with our earlier findings (Holbrook, Cho, and Johnson 2006). More sensitive questions and questions with showcards also showed more mapping problems. The latter finding is somewhat surprising since showcards are designed to aid in the mapping step, but showcards may also be more likely for response formats that are difficult to convey (e.g., semantic differentials with only the endpoints labeled or categorical questions with a large number of response options). Finally, as with comprehension problems, respondent mapping problems also reduced over the course of the interview, suggesting that respondents learn over the course of the interview. A model predicting respondent mapping difficulties that also includes interviewer reading errors as an independent variable is shown in model 5. Interviewer reading errors were unassociated with respondent mapping problems.

Interviewer probing errors are predicted in model 6. Somewhat unexpectedly, deliberately difficult questions were associated with fewer probing errors, perhaps because the purpose of these questions and interviewer responses to problems with them was extensively covered in interviewer training (i.e., it is unusual for interviewers to be given questions that are designed to be problematic and so this was addressed at length in interviewer-training and mock interviews). One might expect question characteristics to influence interviewer probing errors similarly to respondent mapping difficulties, because the former is often the reason why probing is necessary. Consistent with this, the effect of response format on interviewer probing errors was similar to its effect on respondent mapping problems, as were the effects of abstraction, sensitivity, and requesting a time-qualified judgment. In contrast, interviewer probing problems were more likely for deliberately difficult questions and for proxy questions – question characteristics that were not associated with respondent mapping difficulties.

When respondent comprehension and mapping problems were added to the model predicting interviewer probing problems in model 7, both were significant positive predictors. Interviewers were less likely to probe appropriately when the respondent demonstrated either a comprehension or mapping problem. Furthermore, there is weak evidence that respondent problems may partly account for the impact of question characteristics on interviewer probing problems. In particular, the effect of the set of dummy variables representing response formats was smaller in the model shown in equation 7 than the model in model 6, although the effects of other predictors remained virtually unchanged (e.g., abstraction) or were even strengthened (e.g., whether or not the question was a proxy report).

Implications, Limitations, and Next Steps

Our findings replicate previous evidence about the effects of question characteristics on respondent difficulties (e.g., Holbrook, Cho, and Johnson 2006), specifically evidence that text reading level and response format are important predictors. We did not replicate the finding that more abstract questions resulted in more respondent difficulties and found that time qualified judgments led only to more comprehension difficulties (not more mapping difficulties as we found in our previous research). However, we note that this research examined a much broader array of question characteristics (for a larger number of questions) and that some of these characteristics were more systematically manipulated than in our past research. Although this is a strength of the current research, the number of question characteristics examined may raise potential concerns about testing them all simultaneously. (Although many of the factors we examined were manipulated independently and we found relatively weak associations even between question characteristics that were measured.)

Another extension of this research was that we considered interviewer behaviors. We found that question characteristics predicted interviewer behaviors – particularly the type of response format used in the question. However, we found no evidence that interviewer reading errors led to either respondent comprehension or mapping problems, but we did find that respondent comprehension and mapping problems were predictors of interviewer probing problems. These findings are consistent with other evidence examining respondent-interviewer interactions that problematic sequences may be more likely to be initiated by a problematic respondent behavior than by a problematic interviewer behavior (Ongena 2005).

One of the limitations of our research was that respondents were not randomly assigned to interviewers, because interviewers were race/ethnicity matched to respondents. This is a limitation because we cannot fully consider the effects of interviewer characteristics, but it may also be considered a strength because we wanted to hold the race/ethnicity match/mismatch dimension constant. A second limitation is that the sample used in our study was not a probability sample (and was stratified by race/ethnicity and language) and therefore is not representative of a population.

One of the strengths of these data are that they are stratified by race/ethnicity and language and future analyses of these data will focus on potential interactions between question characteristics and respondent characteristics. For example, question characteristics may have different effects on question comprehension among respondents from different racial or ethnic groups or among respondents from the same racial and ethnic group with different levels of acculturation.

Acknowledgement

Some of the analyses in this paper were reported in a paper presented at the 2015 annual conference of the American Association for Public Opinion Research. This work was supported by grants from the National Science Foundation [0648539 to A.L.H., T.P.J., Y.I.C., S.S., N.C., and S.W.]; and the National Institutes of Health [1RO1HD053636-01 to T.P.J., A.L.H., Y.I.C., S.S., N.C., and S.W.].

REFERENCES

- Dykema, J., J.M. Lepkowski, and S. Blixt. 1997. "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley.
- Holbrook, A.L., S. Anand, T.P. Johnson, Y.I. Cho, S. Shavitt, N. Chavez, and S. Weiner. 2014. "Response Heaping in Interviewer-Administered Surveys: Is It Really a Form of Satisficing?" *Public Opinion Quarterly* 78 (3): 591–633.
- Holbrook, A.L., Y.I. Cho, and T.P. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70 (4): 565–95.
- Ongena, Y.P. 2005. "Interviewer and Respondent Interaction in Survey Interviews." <http://dspace.ubvu.vu.nl/bitstream/handle/1871/9842/6353.pdf;jsessionid=2E5D81F671AF969611468392364315CB?sequence=1>.
- Willis, G. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.
- Zouwen, J. van der, W. Dijkstra, and J.H. Smit. 1991. "Studying Respondent-Interviewer Interaction: The Relationship between Interviewing Style, Interviewer Behavior, and Response Behavior." In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman. New York: Wiley.