Survey Practice

Vol. 9, no Spl. Issu., 2016 | www.surveypractice.org

The premier e-journal resource for the public opinion and survey research community

The Impact of Parenthetical Phrases on Interviewers' and Respondents' Processing of Survey Questions

Jennifer Dykema
University of Wisconsin Survey Center
University of Wisconsin-Madison

Nora Cate Schaeffer
Department of Sociology
University of Wisconsin Survey Center
University of Wisconsin-Madison

Dana Garbarski
Department of Sociology
Loyola University Chicago

Erik V. Nordheim
Department of Statistics
University of Wisconsin-Madison

Mark Banghart
Social Science Computing Cooperative
University of Wisconsin-Madison

Kristen Cyffka
Max Planck Institute for Demographic Research

Abstract

Many surveys contain sets of questions (e.g., batteries), in which the same phrase, such as a reference period or a set of response categories, applies across the set. When formatting questions for interviewer administration, question writers often enclose these repeated phrases in parentheses to signal that interviewers have the option of reading the phrase. Little research, however, examines what impact this practice has on data quality. We explore whether the presence and use of parenthetical statements is associated with indicators of processing

Publisher: AAPOR (American Association for Public Opinion Research)
Suggested Citation: Dykema J., N. C. Schaeffer, D. Garbarski, E. V. Nordheim, M. Banghart and K. Cyffka. 2016. The Impact of Parenthetical Phrases on Interviewers' and Respondents' Processing of Survey Questions.
Special Issue: Survey Research & Methodology Training
Survey Practice. 9 (Spl. issu.). ISSN: 2168-0094

problems for both interviewers and respondents, including the interviewer's ability to read the question exactly as worded, and the respondent's ability to answer the question without displaying problems answering (e.g., expressing uncertainty). Data are from questions about physical and mental health from 355 digitally recorded, transcribed, and interaction-coded telephone interviews. We implement a mixed-effects model with crossed random effects and nested and crossed fixed effects. The models also control for some respondent and interviewer characteristics. Findings indicate respondents are less likely to exhibit a problem when parentheticals are read, but reading the parentheticals increase the odds (marginally significant) that interviewers will make a reading error.

Introduction

Question writers often enclose phrases that are repeated from an earlier question in parentheses to signal that interviewers have the option of reading or omitting the phrase. For example, the following questions appeared in the 2003–2005 telephone interview for the Wisconsin Longitudinal Study (WLS) without accompanying instructions for interviewers regarding when to read the repeated phrases: "(During the past four weeks) Have people who do not know you understood you completely when you speak?" and "(Has a doctor ever told you that you have) Cancer or a malignant tumor not including minor skin cancers?"

Optional phrases are commonly used, but little research examines what impact, if any, they have on data quality. In our literature review, we uncovered only a single study that directly examined the effect of parenthetical phrases on data quality. In their analysis exploring the joint effects of question, respondent, and interviewer characteristics on administration times in a telephone interview, Olson and Smyth (2015) reported none of the visual design features they examined, including whether their questions featured parenthetical statements, had any impact on response time. In contrast, the authors found several other question characteristics – such as length, reading level, and the number and format of response options – were associated with longer administration times.

Olson and Smyth's (2015) study, like ours, can be located in an emerging body of research within questionnaire design focusing on analysis of question characteristics provided by nonexperimental or "observational approaches." With observational approaches, researchers identify a set of individual item characteristics (e.g., response format, question length, question difficulty); code questions based on how they vary across the characteristics; and examine their relationship with an outcome related to data quality, such as response latencies (Schaeffer and Dykema 2011). Characteristics analyzed in an observational study typically are chosen using an ad hoc approach, such as to meet goals of a particular analysis (e.g., Holbrook et al. 2006; Yan and Tourangeau 2008), or using a system-based approach in which researchers identify problematic question characteristics by associating "problems" interviewers or respondents

encounter with specific features of survey questions (e.g., Saris and Gallhofer 2007). Observational approaches differ from more traditional experimental approaches where researchers select questions with specific characteristics, vary only one or two features of interest (e.g., number of response categories) while holding other characteristics of the question constant, and examine those features on a criterion built into the experiment's design.

In the current study, we examine the relationship between the inclusion and use of parenthetical phrases and the likelihood that interviewers and respondents will exhibit indictors of problematic interactional behaviors (i.e., which could be associated with measurement error). Like other observational studies, our models control for additional question characteristics and select characteristics of respondents and interviewers.

Methods

Survey Data

Data are provided by the 2003–2005 telephone administration of the WLS, a longitudinal study of a one-third random sample of the 1957 class of Wisconsin high school graduates (n=10,317) (Sewell et al. 2003; AAPOR RR2=80 percent). Our analytic sample of 355 cases was randomly selected as follows (see Garbarski et al. [2011] for further details). The WLS sample was divided into random replicates. We used even-numbered replicates to distribute sample over the field period, and randomly selected 100 interviewers from the 137 completing 4 or more interviews during even-numbered replicates. Due to budget constraints, we selected between three and five respondents from each of the sampled interviewers. To sample respondents within interviewers, we stratified by the respondent's cognitive ability – assessed by their high school IQ score and normalized for the sample – and randomly selected two respondents with low and high cognitive ability, and one with medium cognitive ability. Stratifying by IQ ensured we had an adequate number of respondents with lower measures of cognitive ability to examine variation across ability levels. Comparisons between our analytic sample and the entire sample indicated they were similar across a range of sociodemographic characteristics.

We examine interviewer-respondent interaction during a series of health questions. These questions were the first substantive module in the survey and contained items about self-rated health, physical and mental health functioning, and diagnosed health conditions (http://www.ssc.wisc.edu/wlsresearch/documentation/flowcharts/Full_Instrumentation_1957_2010_vers8_Final.pdf). Although many of the questions are from standardized instruments, such as the Health Utilities Index, prior qualitative work demonstrated they were written in ways difficult for older respondents to process. While 76 questions were included in this module, respondents received fewer questions because of skip patterns, and we limited analysis to the 23 questions administered to all respondents.

Interaction Coding

We identified over 100 behaviors for coding based on a conversation analysis of a subset of the transcripts, detailed examination of the interviews used for the conversation analysis, and the literature on interaction in survey interviews. Coding was done from transcripts using the Sequence Viewer program (Wil Dijkstra, http://www.sequenceviewer.nl/) by five former WLS interviewers, who received extensive training. To assess intercoder reliability, a sample of 30 cases was independently double-coded by five coders, and a measure of inter-rater agreement, Cohen's Kappa, was produced. While Kappa values varied across the behaviors coded (available upon request), the average overall Kappa for all coded events in the health section was high at 0.861. Our unit of analysis is the question-answer sequence (n=8,150), which begins with the interviewer's question reading and ends with the last utterance spoken by the interviewer or respondent before the interviewer reads the next question.

Measures

Table 1 shows descriptive statistics for the variables in the analysis. Two binary outcomes serve as indicators of problems with processing and are associated with measurement error. First, we assess how accurately interviewers read the questions. We code readings as exact versus any change. Changes included slight change (i.e., diverges slightly from the script but does not change the meaning of the script), major change (i.e., diverges from the script in a way that changes the meaning), or verification (i.e., alters wording of initial question-asking to take into account information provided earlier). Second, for respondents, we look at an index of behaviors indicative of potential problems answering the question, including providing reports, considerations, expressions of uncertainty, and other uncodable answers. For the analysis, the index is collapsed to a binary indicator of no problems versus one or more problems.

Questions are classified based on their values for the following characteristics: parenthetical administration, response format, and question length. Questions are coded with regard to whether they did not include a parenthetical phase, included a parenthetical phrase that was not read, or included a parenthetical phrase that was read. Response format refers to how the question is formatted for response. Two formats appear in our data: yes-no questions that provide "yes" or "no" as categories and selection questions that provide a set of predetermined categories. Question length is measured as the raw number of words in the question. For questions including parenthetical phrases, we made separate assessments of word counts with and without the parenthetical using information from coders about whether the interviewer included the phrase.

We include several characteristics of respondents. Gender is coded 1 if male and 0 if female. Education is measured in years of schooling. Cognitive ability is indicated by the respondent's IQ score, assessed during the respondents' freshman and junior years of high school using the Henmon-Nelson test of mental ability.

Table 1 Descriptive statistics for behavioral outcomes, question characteristics, respondent characteristics, and interviewer characteristics.

	Mean or percent	SD	Minimum	Maximum	n
Dependent variables[a]					
Interviewer exact question reading	71.03		0	1	8,150
Any respondent problem behaviors	25.23		0	1	8,150
Independent variables[b]					
Question characteristics					
Parenthetical administration					
No parenthetical in question	43.48				10
Parenthetical in question read	43.48				10
Parenthetical in question not read	13.04				3
Response format					
Yes/no	82.61				19
Selection	17.39				4
Question length (in words)	18.47	10.24	1	43	23
Respondent characteristics[c]					
Gender (male)	47.89		0	1	355
Education (in years)	13.68	2.37	12	20	355
Cognitive ability	101.97	17.98	67	145	355
Health status					
Bottom tertile	35.77				127
Middle tertile	26.48				94
Top tertile	30.42				108
Missing	7.32				26
Interviewer characteristics[d]					
Experience (in months)	13.80	16.55	0	67	79

[a]Column labeled "n" shows the total number of question-answer sequences in the analysis.
[b]Column labeled "n" shows the number of questions in total and in each category.
[c]Column labeled "n" shows the number of respondents in total and each category.
[d]Column labeled "n" shows the number of interviewers.

Graduates who participated in the telephone interview were subsequently sent a mailed self-administered questionnaire (SAQ), which we use to obtain a measure of the respondent's health status (using the physical component summary from the Short Form Health Survey [SF-12]; Ware et al. 1996) that is exogenous to the health questions administered during the telephone survey. We collapse values from the SF-12 into tertiles for respondents who completed the SAQ. For interviewers, we examine the effect of experience, measured as the numbers of months of interviewing experience prior to the telephone survey.

Analysis

To account for the complicated crossed and nested structure of the data, we implement a mixed-effects model with a variance structure that uses crossed random effects. Initial models included random effects for interviewers, questions, and respondents (nested within interviewers and crossed with

question). However, results indicated that including all three random effects resulted in the models being overfitted; removing the respondent random effect reduced the overfit. Question, respondent, and interviewer characteristics are modeled as fixed effects which are nested within and crossed with the random effects. The response variables are binary; logit models were computed in R using the glmer function from the lme4 package.

Results

Results are presented in Table 2 and shown separately for the outcomes of exact question reading by interviews and any problem behaviors by respondents. The odds ratio provides the proportional change in the odds of interviewers reading exactly or respondents exhibiting a problem behavior. The first section of results is for question characteristics. When questions include parenthetical phrases, and interviewers read them, the odds of reading the question exactly are lower (marginally significant, $p<0.10$) compared to the odds when questions contain

Table 2 Multilevel logistic regression analyses of interviewer exact question reading and respondent problem behaviors on question, respondent, and interviewer characteristics.

	Interviewer exact question reading		Any respondent problem behaviors	
	Odds ratio	95% Confidence interval	Odds ratio	95% Confidence interval
Question characteristics				
Parenthetical administration				
[Parenthetical in question not read]	–		–	
No parenthetical in question	0.87	[–0.61, 0.36]	1.19	[–0.60, 0.94]
Parenthetical in question read	0.77+	[–0.57, 0.05]	0.63*	[–0.83, –1.60]
Selection response format [vs. yes/no]	1.63	[–0.17, 1.17]	3.38*	[0.16, 2.31]
Question length (in words)	0.91***	[–0.12, –0.07]	1.07**	[0.03, 0.04]
Respondent characteristics				
Male (vs. female)	0.97	[–0.17, 0.11]	0.98	[–0.15, 0.12]
Education (in years)	0.96*	[–0.08, –0.01]	1.01	[–0.03, 0.04]
Cognitive ability	1.00	[–0.00, 0.01]	0.99**	[–0.01, –0.00]
Health status				
Bottom tertile	0.77	[–0.47, –0.06]	1.63***	[0.30, 0.68]
Middle tertile	0.92	[–0.26, 0.08]	1.26**	[0.07, 0.39]
[Top tertile]	–		–	
Missing	0.82	[–0.47, 0.08]	1.73***	[0.29, 0.81]
Interviewer characteristics				
Experience (in months)	1.01	[–0.02, 0.03]	1.00	[–0.01, 0.00]
Intercept	36.41***	[2.78, 4.43]	0.08***	[–3.39, –1.60]

+$p<0.10$, *$p<0.05$, **$p<0.01$, ***$p<0.001$.

parenthetical statements, but they are not read. For respondents, hearing the parenthetical phrase, when one is available, is associated with lower odds of exhibiting a problem behavior ($p<0.05$) relative to when the parenthetical phrase is not read. There are no significant differences in the odds of interviewers reading a question exactly or respondents exhibiting problems answering when questions do not contain parenthetical phrases relative to when questions contain unread parenthetical phrases. Whether the question is formatted for a yes/no response or selection from predetermined categories does not affect interviewers' ability to deliver the question as worded, but selection questions are associated with higher odds that respondents exhibit a problem. Question length has a negative impact on both interviewers and respondents: Longer questions are significantly associated with lower odds of interviewers reading questions exactly and higher odds of respondents having problems answering. (However, note that for interviewer question-reading, the odds ratio of 0.91 for question length is significant whereas the odds ratio of 1.63 for a selection format is not. This is because question length, measured as a continuous variable, contains more information than the binary response format variable).

Gender is not associated with either of the outcomes examined. However, interviewers appear to have more difficulty reading questions exactly as worded when interacting with respondents with lower education levels and those whose self-assessed physical health status is poorest (i.e., in the bottom tertile of the SF-12). There are negative associations between the likelihood of displaying a problem behavior and respondents' cognitive ability and health: Respondents with lower cognitive ability, with physical health status ratings in the bottom or middle tertiles, or who failed to respond to the health questions in the SAQ each have higher odds of exhibiting problems. Interviewers' experience is unrelated to the outcomes.

Discussion

In contrast to previous research which showed inclusion of parenthetical phrases was not related to response times in a telephone interview (Olson and Smyth 2015), in this study the inclusion of parenthetical phrases has implications for question reading and question answering. When questions include parenthetical statements and they are read, respondents are less likely to demonstrate problems answering the question than if they are not read. This suggests two possibilities: (1) interviewers might effectively use information from their preceding interaction with respondents to make decisions about when to include the parenthetical, or (2) respondents might benefit from the parenthetical text whether or not they earlier signaled problems to the interviewer. However, when interviewers include parenthetical information there is a marginally significant association with lower odds of interviewers reading the questions exactly as worded, possibly because inclusion of the parenthetical adds extra words and more opportunities to make a reading error or because attempting to include a parenthetical on the spur of the moment leads to reading mistakes. Overall, our findings point to a potential

tradeoff in the inclusion of parenthetical phrases for standardized measurement: When interviewers include parenthetical phrases, respondents' answers may be of higher quality; however, inclusion of these phrases – which are optional and used at interviewers' discretion – may increase interviewer variability. We add that the "feedback loop," relating interviewer inclusion of parentheticals to prior respondent behaviors – which consequently might affect interviewer error, requires us to interpret the results cautiously.

Consistent with prior research, our findings regarding question length show that longer questions are associated with negative outcomes – questions are more likely to be read in error and respondents are less likely to provide an immediately codable answer. We caution the reader not to interpret these findings as suggesting longer questions will necessarily yield poorer quality data; the questions we analyzed that were long were also complex in other ways, and interviewers who read the parenthetical were implementing the decision to do so on the fly. Our findings do, however, suggest support for the common recommendation to write questions as simply as possible.

Respondents' characteristics also showed interesting associations with interviewers' question reading and respondents' processing problems. Respondents with fewer years of education and in poorer health may exhibit interactional cues (e.g., pausing longer before answering) in prior interactions that signal comprehension problems and subsequently increase the likelihood interviewers will add a parenthetical phrase to make a question clearer. Thus, we suggest future research explore factors that predict interviewers' use of parenthetical statements. This research should also examine patterns of interactional behaviors across questions. Consistent with prior research, we find respondents' cognitive ability and health status are associated with respondents having processing problems in expected directions. Future research should also examine interactions between characteristics of respondents – such as cognitive ability – and interviewers' use of parenthetical phrases in predicting interviewers' question reading and respondents' processing problems.

We note several limitations with our research. First, like many observational studies, we lack direct measures of validity and reliability to assess the impact of question, respondent, and interviewer characteristics. Instead, we use question-asking and question-answering behaviors as proxy measures of measurement error. Second, our questions are not randomly sampled from a population of questions with many different characteristics. This highlights an advantage of observational studies: they feature items administered in an actual operational setting, but related disadvantages are that the items may have a limited range of characteristics (e.g., our questions had limited response formats) and might not conform to current best practices. Third, our study is one of a few observational studies that examine the effects of question, respondent, and interviewer characteristics using binary outcomes. In contrast to continuous outcomes (e.g., response times), less information is available for analysis and small group sizes resulting from combinations of characteristics can cause difficulties estimating mixed-effects models. For example, our models could not provide reliable and defensible results if we included all three of the desired random effects. Based

on extensive assessment of our models (including examination of the stability of our parameter estimates when variables were added or removed), we feel more confident (statistically) in our results for respondent processing problems than interviewer question reading. High levels of multicollinearity among predictor variables in conjunction with a small number of data points among combinations of predictor variables contribute to the instability. Also, we were unable to include the random error term for respondents (to avoid overfitting the model), and our results could potentially be altered by including this term.

Author Note

This research was supported by the following at the University of Wisconsin-Madison: a National Institute on Aging grant to Schaeffer (under P01 AG 21079 to Robert M. Hauser); the Center for the Demography of Health and Aging (NIA Center Grant P30 AG017266); the University of Wisconsin Survey Center (UWSC); and the use of facilities of the Social Science Computing Cooperative and the Center for Demography and Ecology (NICHD core grant P2C HD047873). This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, AG-033285, and AG-041868), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison. Since 1992, data have been collected by the University of Wisconsin Survey Center. A public use file of data from the Wisconsin Longitudinal Study is available from the Wisconsin Longitudinal Study, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 and at http://www.ssc.wisc.edu/wlsresearch/data/. The opinions expressed herein are those of the authors.

References

Garbarski, D., N.C. Schaeffer and J. Dykema. 2011. Are interactional behaviors exhibited when the self-reported health question is asked associated with health status? *Social Science Research 40*(4): 1025–1036.

Holbrook, A., Y.I. Cho and T. Johnson. 2006. The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly 70*(4): 565–595.

Olson, K. and J.D. Smyth. 2015. The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*. 3(3): 361–396.

Saris, W.E. and I.N. Gallhofer. 2007. *Design, evaluation, and analysis of questionnaires for survey research*. Wiley, New York.

Schaeffer, N.C. and J. Dykema. 2011. Questions for surveys: current trends and future directions. *Public Opinion Quarterly 75*(5): 909–961.

Sewell, W.H., R.M. Hauser, K.W. Springer and T.S. Hauser. 2003. As we age: a review of the Wisconsin Longitudinal Study, 1957–2001. Research in Social Stratification and Mobility 20: 3–111.

Ware, J.E. Jr., M. Kosinski and S.D. Keller. 1996. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical Care 34*(3): 220–233.

Yan, T. and R. Tourangeau. 2008. Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology 22*(1): 51–68.