View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DigitalCommons@University of Nebraska

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

2019 Workshop: Interviewers and Their Effects from a Total Survey Error Perspective

Sociology, Department of

2020

Chapter 18: Response Times as an Indicator of Data Quality: Associations with Question, Interviewer, and Respondent Characteristics in a Health Survey of Diverse Respondents. Appendix 18

Dana Garbarski

Jennifer Dykema

Nora Cate Schaeffer

Dorothy Farrar Edwards

Follow this and additional works at: https://digitalcommons.unl.edu/sociw

Part of the Quantitative, Qualitative, Comparative, and Historical Methodologies Commons

This Article is brought to you for free and open access by the Sociology, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in 2019 Workshop: Interviewers and Their Effects from a Total Survey Error Perspective by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Chapter 18: Response Times as an Indicator of Data Quality: Associations with Question, Interviewer, and Respondent Characteristics in a Health Survey of Diverse Respondents

Appendix 18

Dana Garbarski, Jennifer Dykema, Nora Cate Schaeffer, and Dorothy Farrar Edwards

Full book reference:

Olson, K., J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, B. T. West. 2020. *Interviewer Effects from a Total Survey Error Perspective*. Boca Raton: CRC Press.

Appendix 18A Description of individual question characteristics and hypotheses for their relationship with RTs

We examine whether individual question characteristics available in our survey are associated with response times (RTs) based on relationships demonstrated in previous research and our expectations about whether the question characteristic is likely to increase the cognitive processing burden for the respondent, interviewer, or both. Some of hypotheses are evident; some require explication. All hypotheses are made under the assumption that other question characteristics are held constant.

Hypothesis 1a. Number of words. Because the number of words increases question-asking time and possibly question-answering time, we predict word count will be associated with longer RTs (Couper and Kreuter 2013; Olson and Smyth 2015; Yan and Tourangeau 2008).

Hypothesis 1b. Question order. We predict that questions asked later in the interview will have shorter RTs than questions asked earlier in the interview because 1) respondents learn the process of answering survey questions as the interview unfolds, 2) interviewers increase their pace to get through longer surveys more quickly, 3) respondents satisfice in their responses as the interview progresses, and 4) respondents learn the sort of answer that is needed (Garbarski et al. 2016; Holbrook et al. Chapter 17; Krosnick 1991).

Hypothesis 1c. Question type. We predict personal sociodemographic questions will have the shortest RTs because the information required to answer these questions is readily accessible and these types of questions are commonly asked of and answered by respondents (Olson and Smyth

2015; Yan and Tourangeau 2008). Questions about respondents' behaviors or specific events (factual questions) will require more time for memory retrieval than demographic questions. Questions about subjective assessments or attitudes will take the longest time to answer, particularly if the respondent does not have a preformed attitude readily accessible (Tourangeau, Rips and Rasinski 2000).

Hypothesis 1d. Question form. We predict yes/no questions will be associated with shorter RTs compared to other question forms because the yes/no response options are implied by the format of the question and not read to the respondent; the only cognitive processing required is deciding between two response options. We expect the other question forms in our survey will be associated with longer RTs compared to yes/no questions. Ordinal question forms – either unipolar or bipolar – present respondents with ordered response categories and require that the respondent consider all of the categories offered because the meaning of a single category (e.g., "very") depends on the entire set of categories (Schaeffer and Charng 1991). Nominal questions require interviewers read and respondents consider all the choices in a list of response options (Olson and Smyth 2015). Open question formats lead to open, varied, and on average longer answers (Couper and Kreuter 2013), although this depends on the type of open question. For example, discrete-value questions are a type of open format that requires a numerical value as an answer; these answers may not differ from yes/no questions if the information asked for is salient and accessible.

Question characteristics may also impact RTs because they increase the complexity of the task for the actors in the survey interview (Olson and Smyth 2015). Some features may have a greater influence on complexity for respondents than interviewers (or vice-versa), and "complexity" itself likely differs for each actor given their different roles and goals in the task (Garbarski et al. 2016). We use the following framework to discuss question characteristics that we expect will influence the complexity or efficiency of the task for 1) respondents or 2) interviewers.

Respondents' task complexity

Hypothesis 1e. Definition in the question. Questions often contain definitions. Definitions may increase RTs by adding words to a question (Olson and Smyth 2015). Controlling for other question characteristics such as number of words or being administered as a parenthetical phrase (that is, at the interviewer's discretion), definitions are also more common with complex questions (such that a definition is needed) and may be associated with longer RTs.

Hypothesis 1f. List-item question. "List item" questions are those that list a set of objects linked with the words "or" or "and" (Dykema et al. 2009). We predicted that questions featuring a list-item format lead to longer RTs because their many parts require extra time for interviewers to read and extra time for respondents to process.

Hypothesis 1g. Sensitive question. Respondents may find questions about sensitive topics intrusive, threatening, or emotionally painful. They may not want to report sensitive information out of concern for appearance or repercussions such as when questions ask about illegal behaviors (Olson and Smyth 2015; Schaeffer 2000; Tourangeau and Yan 2007). Providing a response to a sensitive question may lead to longer RTs if respondents hesitate or modify their answer. Alternatively, however, respondents may speed through the interaction in response to a sensitive question. So might interviewers, who, under the constraints of standardization, are not able to expand on the meanings and definitions of particular concepts (e.g., "so would you say yes then?") (Garbarski et al. 2016). Thus, we predict that sensitive questions will be associated with shorter RTs.

Hypothesis 1h. Race-related question. Questions can also be complex for respondents because they ask about topics respondents have not considered or not considered within the manner being asked (Bassili and Fletcher 1991). In this study, the topic of race/ethnicity is a factor we expect to influence the complexity of the task for respondents in a way that will increase overall RTs (Dykema et al. forthcoming). The survey discusses barriers and facilitators of medical research, and at several points asks respondents to consider the treatment of their particular racial/ethnic group beyond their personal experience, something that we expect to be a task for which pre-formed attitudes are not readily accessible. (These questions are not necessarily sensitive; only two of the nine questions that mentioned race were also coded as sensitive.) Thus, we predict that questions about how one's own racial/ethnic group is treated within the context of medical research may be unfamiliar to respondents and associated with longer RTs.

Hypothesis 1i. Battery structure. Question structure refers to the relationship between questions (Schaeffer and Dykema 2015). According to Alwin and Beattie (2016), a battery refers to adjacent questions on the same topic with the same set of response options. A series refers to adjacent questions on the same topic that do not have the same set of response options. Stand-alone questions are those not topically related to adjacent questions. Here we focus on whether a question appears in a battery, series, or stands alone; and if part of a battery or series, whether it appears first or later. Given how questions are linked across topics, respondents should be able to answer the later questions in a series or battery more quickly because they have tapped into the memory structure available for the broader topic (Tourangeau et al. 2000) and, for batteries, are familiar with that set of response options (Olson and Smyth 2015). Olson, Smyth, and Cochran (2018) posit that batteries constitute an opportunity for learning the pattern of the set of questions. They found evidence of more interviewer–respondent interaction, higher rates of including the question stem and response options in initial question reading, and probing with the question stem or response options for items asked earlier in the battery compared to those asked later. Questions that are part of batteries contain features of other question characteristics described here. However, we still expect an independent effect of battery structure holding constant the number of words in a question (because we expect that questions that are first in a battery or a series would have more words to introduce the topic of the set of questions), emphasis (described below), and parenthetical phrases (described below). Thus, we expect that questions later in a battery (or series) will have shorter RTs than those that occur earlier in the battery (or series) because respondents learn the question topic (for both batteries and series) and question structure (for batteries).

Hypothesis 1j. Emphasis in the question. Emphasis in a question's wording (e.g., placing words in bolded, underlined, or italicized text) may increase or decrease RTs. Olson and Smyth (2015) hypothesized emphasis would increase RTs by increasing the complexity of the interviewer's task. However, emphasis could decrease RTs if this practice made the respondent's processing more efficient by clarifying meaning in the question (e.g., by making a contrast with previous questions) or providing a threshold for answering (e.g., "any" concern). Emphasis is often used in questions that are part of a battery to help distinguish how the focal question differs from the preceding questions. Even controlling for the battery structure, however, emphasis should have an independent effect on RT because emphasis helps to distinguish what is unique about the current question from prior questions within the series or battery. Thus, we predict that questions with emphasis will have shorter RTs because emphasis serves to facilitate respondents' processing by clarifying the meaning or threshold of the question.

Interviewers' task complexity

The survey includes two question characteristics – interviewer instructions and parenthetical phrases – that may increase the complexity of the interviewer's task and RTs as a result.

Hypothesis 1k. Interviewer instructions. Instructions to interviewers (e.g., "How many years have you lived in Wisconsin? (INTERVIEWER: ROUND UP): 1 year or less, 2 to 10 years, 11 to 25 years, More than 25 years, IF VOLUNTEERED: All my life") may increase the complexity of the task by giving interviewers more to read, process, and act on. Thus, we predict that interviewer instructions will be associated with longer RTs. While Olson and Smyth (2015) found no relationship between interviewer instructions and RTs, Couper and Kreuter (2013) found that interview instructions were associated with shorter RTs. We return to this relationship in our expectations for interactions with interviewers' experience.

Hypothesis 1l. Parenthetical phrases. Phrases that are repeated from an earlier question in a battery or series may be placed in parentheses to signal to interviewers that reading the text is optional—e.g., "(Has a doctor ever told you that you have) Alzheimer's disease or dementia?"

when reading a list of health conditions. Parenthetical phrases have no association with RTs in one study (Olson and Smyth 2015) and are also associated with increased odds of inexact and disfluent question reading (Dykema et al. 2016; Olson et al. 2019). When interviewers do read parenthetical phrases, respondents appear to have fewer problems answering questions (Dykema et al. 2016)—so more words read by interviewers, but less disfluency in answers from respondents.[1] On balance, we expect that interviewers consider reading parenthetical phrases as optional, leading to our prediction that parenthetical phrases will be associated with shorter RTs.

[1] 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 their earlier behavior signaled problems to the interviewer.

Appendix 18B Description of established tools for evaluating questions and hypotheses for their relationship with RTs

Hypothesis 2a. Flesch-Kincaid grade level. The Flesch-Kincaid grade level score is based on the average number of words per sentence and the average number of syllables per word, with higher scores indicating that the text requires a higher grade level to understand. However, its utility as a measure of survey question difficulty is questionable given that questions tend to have short text, inconsistent punctuation, and rely on conversational practices to communicate meaning (Dykema et al. 2019; Lenzner 2014). Olson and Smyth (2015) reported that questions with higher reading levels (harder to read) took longer to administer. Thus, we predict grade level scores will be positively associated with RTs.

Hypothesis 2b. QUAID. The Question Understanding Aid (QUAID) is an online program in which users enter a question's text to receive a list of problems related to comprehension difficulty, such as whether the question contained unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun phrases, complex syntax, or working memory overload (Graesser et al. 2006; http://quaid.cohmetrix.com/). We predict an indicator summarizing problems identified by QUAID will be positively associated with RTs.

Hypothesis 2c. QAS. The Question Appraisal System (QAS) is a coding scheme intended to identify 27 problems a question could present for respondents or interviewers based on a question's characteristics in the following broad categories: reading, instructions, clarity, assumptions, knowledge/memory, sensitivity bias, and response categories (Willis 2005; Willis and Lessler 1999). For example, a response category characteristic coded in the QAS is whether there is a mismatch between the question and response categories. We predict an indicator totalling the number of problems identified by QAS will be positively associated with RTs.

Hypothesis 2d. QAS. The Survey Quality Predictor (SQP) is a comprehensive online tool for coding up to 50 different question characteristics related to features of the language, structure, content, and administration of a question (Saris and Gallhofer 2007; http://sqp.upf.edu/). After an item is coded by human coder, the online tool produces predicted reliability, validity, and quality (the product of reliability and validity) estimates using coefficients that were previously estimated in empirical analyses of split-ballot multi-trait multi-method (MTMM) experiments. The quality of the predictions depend on the specification of the original model, the adequacy of the original data, and the accuracy of the coding. We predict the quality indictor produced by SQP, in which a higher score indicates higher quality, will be negatively associated with RTs.

Appendix 18C Sample Description

Volunteer list. For the volunteer sample, members of the project team recruited 471 (n = 46 White, n = 137 Black, n = 144 Latino, and n = 144 American Indian) individuals through connections they built with leaders in specific racial and ethnic communities, by visiting churches and community centers, by attending events sponsored by specific racial or ethnic groups (e.g., pow-wows), and by posting flyers at targeted locations in communities. Project staff collected names, demographic data (e.g., race and ethnicity), and contact information (e.g., phone numbers) for these potential respondents, and all individuals identified through these channels were contacted and asked to participate in the study.

Vendor list. A total of 8,075 records were purchased from Infogroup, a business and consumer data provider. Infogroup filtered data from their databases based on a surname algorithm and geo-coding that would supposedly help target individuals living in diverse communities in Wisconsin. In addition, Infogroup filtered records to accrue only those with high-deliverability for direct mail and those with active telephone numbers. From the list of records, a total of 700 cases (7 replicates of 100 cases each, consisting overall of 100 White, 200 Black, 200 Latino, and 200 American Indian targeted individuals) were fielded for calling.

Table C1 shows the distribution of completed interviews by the respondent's race/ethnicity for the volunteer and vendor lists.

Table 18.C1. Number of completed interviews by respondents' race/ethnicity and sample types

Race/Ethnicity	Volunteer List	Vendor List
White	29	73
Black	103	3
Latino	93	7
American Indian	101	1
Total	326	84

Appendix 18D Additional Tables

Table 18D.1 Regression of response times on characteristics of respondents and interviewers, Voices Heard Study

Variables	Mean or Percent		Std. Dev.	Min.	Max.	Coef.	Std. Err.	
Respondents' race/ethnicity								
Black	25.9	%				-0.047	0.017	**
Latino/a	24.4	%				Reference		
American Indian	24.9	%				-0.042	0.017	*
White	24.9	%				-0.083	0.019	***
Respondent is a woman (vs. a man)	65.1	%				0.012	0.013	
Respondents' age (in years)	44.70		16.74	18.00	90.00	0.002	0.000	***
Respondents' education								
High school or less	31.0	%				Reference		
Some college	32.9	%				-0.004	0.015	
College or more	36.1	%				-0.003	0.015	
Interviewers' race is non-white (vs. white)	20.8	%				-0.009	0.040	
Interviewer is a woman (vs. a man)	54.2	%				0.073	0.032	*
Interviewers' age (in years)	24.36		5.56	20.00	42.00	0.004	0.003	
Interviewers' experience is one year or more (vs. less)	37.5	%				-0.012	0.032	
Number of interviews completed	17.08		12.52	1.00	49.00	-0.001	0.001	
Intercept						2.229	0.094	***
Random-effects Parameters								
Interviewer-level variance						0.003	0.001	*
Question-level variance						0.344	0.050	***
Respondent-level variance						0.012	0.001	***
Residual variance						0.084	0.001	***
Wald chi-square						37.03	(df 12)	***
Log restricted-likelihood							-8330.41	

Std. Dev.=Standard deviation; Min.=Minimum, Max.=Maximum, Coef.=Coefficient, Std. Err.=Standard error.
Descriptive statistics are calculated at the level of the respondent (N=410) and interviewer (N=24) for their respective characteristics. Regression analysis is conducted at the level of the question-answer sequence (N=39,052).
*p<0.05, **p<0.01, ***p<0.001

Table 18D.2. Regressions of response times on characteristics of questions, interviewers, and respondents, Voices Heard Study

	Model 1			Model 2		Model 3		Model 4		Model 5	
Variables	Coef.	Std. Err.		Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Number of words	0.020	0.003	***								
Question order	-0.003	0.002	*								
Question type											
Event or behavior (reference category)											
Subjective	0.200	0.151									
Demographic	0.202	0.150									
Question form											
Yes/no (reference category)											
Nominal	0.176	0.125									
Open	0.215	0.188									
Bipolar ordinal	0.985	0.156	***								
Unipolar ordinal	0.585	0.141	***								
Definition in the question (vs. not)	-0.083	0.149									
List-item question(vs. not)	0.078	0.063									
Sensitive question (vs. not)	0.111	0.088									
Race-related question (vs. not)	0.019	0.101									
Battery structure											
First in battery	-0.005	0.111									
Later in battery (reference category)											
First in series	0.252	0.123	*								
Later in series	0.179	0.100									
Standalone	0.195	0.131									

Emphasis in the question (vs. not)	-0.252	0.099	*												
Interviewer instructions (vs. not)	0.225	0.114	*												
Parenthetical phrases (vs. not)	-0.376	0.078	***												
Flesch-Kincaid grade level score				0.059	0.010	***									
QUAID problem score							0.083	0.025	***						
QAS problem score										0.048	0.059				
SQP quality score													1.138	1.253	
Intercept	1.274	0.224	***	1.509	0.151	***	1.867	0.143	***	2.181	0.111	***	1.664	0.629	**
Random-effects Parameters															
Interviewer-level variance	0.003	0.001	*	0.003	0.001	*	0.003	0.001	*	0.003	0.001	*	0.003	0.001	*
Question-level variance	0.048	0.008	***	0.254	0.037	***	0.311	0.045	***	0.346	0.050	***	0.345	0.050	***
Respondent-level variance	0.012	0.001	***	0.012	0.001	***	0.012	0.001	***	0.012	0.001	***	0.012	0.001	***
Residual variance	0.085	0.001	***	0.085	0.001	***	0.085	0.001	***	0.085	0.001	***	0.085	0.001	***
Wald chi-square	635.33	df 31	***	71.5	df 13	***	48.01	df 13	***	37.67	df 13	***	37.85	df 13	***
Log restricted-likelihood		-8264.58			-8319.26			-8327.94			-8332.00			-8328.86	

Coef.=Coefficient, Std. Err.=Standard error.
N=39,052 question-answer sequences
QUAID=Question Understanding Aid, QAS=Question Appraisal, SQP=Survey Quality Predictor
Models control for respondent and interviewer characteristics. Respondent characteristics include race/ethnicity, gender, age, and education. Interviewer characteristics include: race/ethnicity, gender, age, prior interviewing experience, and study-specific experience.
*p<0.05, **p<0.01, ***p<0.001

Table 18D.3 Variance parameter estimates for response times, base model with no covariates and models with covariates

		Interviewer			Question			Respondent			Error			Likelihood Ratio Test	
		Coef.	95% CI		Coef.	95% CI		Coef.	95% CI		Coef.	95% CI			
Base model	Var.	0.004	0.002	0.009	0.344	0.259	0.458	0.013	0.011	0.015	0.085	0.084	0.087	62725.72	****
	VPC	0.010			0.770			0.029			0.191				
Full model	Var.	0.003	0.002	0.008	0.045	0.032	0.062	0.012	0.011	0.014	0.085	0.084	0.087	15995.62	****
	VPC	0.023			0.307			0.084			0.586				
	% red.	21.1%			87.0%			5.3%			0.0%				
R+I	Var.	0.003	0.002	0.008	0.344	0.259	0.458	0.012	0.011	0.014	0.085	0.084	0.087	62496.35	****
	VPC	0.008			0.773			0.028			0.192				
	% red.	21.1%			0.0%			5.3%			0.0%				
Individual question characteristics, R+I	Var.	0.003	0.002	0.008	0.048	0.035	0.066	0.012	0.011	0.014	0.085	0.084	0.087	17367.16	****
	VPC	0.023			0.323			0.082			0.572				
	% red.	21.1%			86.0%			5.3%			0.0%				
Flesch-Kincaid grade level, R+I	Var.	0.003	0.002	0.008	0.254	0.191	0.339	0.012	0.011	0.014	0.085	0.084	0.087	53421.75	****
	VPC	0.010			0.716			0.034			0.240				
	% red.	21.1%			26.1%			5.3%			0.0%				
QUAID problem score, R+I	Var.	0.003	0.002	0.008	0.311	0.234	0.415	0.012	0.011	0.014	0.085	0.084	0.087	59201.09	****
	VPC	0.008			0.755			0.030			0.207				
	% red.	21.1%			9.5%			5.3%			0.0%				
QAS problem score, R+I	Var.	0.003	0.002	0.008	0.346	0.260	0.460	0.012	0.011	0.014	0.085	0.084	0.087	62301.38	****
	VPC	0.008			0.774			0.027			0.191				
	% red.	21.1%			-0.4%			5.3%			0.0%				
SQP problem score, R+I	Var.	0.003	0.002	0.008	0.345	0.259	0.459	0.012	0.011	0.014	0.085	0.084	0.087	62243.19	****
	VPC	0.008			0.773			0.027			0.191				
	% red.	21.1%			-0.2%			5.3%			0.0%				

Var.=variance, VPC=variance partition coefficient, % red.=percent reduced compared to base model.
QUAID=Question Understanding Aid, QAS=Question Appraisal, SQP=Survey Quality Predictor

R+I: Respondent and interviewer characteristics
Full model: Flesch, QUAID, QAS, SQP, individual question characteristics, R+I
N=39,052 question-answer sequences
*p<0.05, **p<0.01, ***p<0.001

Notes. Statistical tests for models and their individual variance components were conducted using likelihood ratio tests (Rabe-Hesketh and Skrondal 2012). We first show the variance partition coefficients (VPCs) for the base model and models that add sets of characteristics, as well as the percentage reduction in variance for a model relative to the base model. The VPCs show the proportion of the variance for each level out of the total variance components (Goldstein, Browne, and Rasbash 2002). As shown by the VPCs for the base model, interviewers, respondents, and questions each contribute a significant portion of the variability in response times, suggesting that a multilevel framework is appropriate for these data.

Appendix **18E References**

Alwin, D. F., and B. A. Beattie. 2016. The KISS Principle in survey design: Question length and data quality. *Sociological methodology* 46:121-152.

Bassili, J. N., and J. F. Fletcher. 1991. RT measurement in survey research: A method for CATI and a new look at nonattitudes. *Public opinion quarterly* 55:331-346.

Couper, M. P. and F. Kreuter. 2013. Using paradata to explore item level RTs in surveys. *Journal of the royal statistical society: Series A (Statistics in society)* 176:271-286.

Dykema, J., D. Garbarski, N. C. Schaeffer, and D. Maynard. 2009. Code manual for codifying interviewer-respondent interaction in surveys of older adults. University of Wisconsin Survey Center, University of Wisconsin-Madison. https://uwsc.wisc.edu/documents/CodeManual_Master_V69_PostProduction.pdf

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. *Survey practice,* 9(2). https://www.surveypractice.org/article/2817-the-impact-of-parenthetical-phrases-on-interviewers-and-respondents-processing-of-survey-questions. Accessed May 31 2019.

Dykema, J., D. Garbarski, N. C. Schaeffer, I. Anadon, and D. F. Edwards. Forthcoming. Correlates of differences in interactional patterns among black and white respondents. In *Understanding survey methodology: sociological theory and application*, ed. P. Brenner. Springer.

Garbarski, D., N.C. Schaeffer, and J. Dykema. 2016. Interviewing practices, conversational practices, and rapport: Responsiveness and engagement in the standardized survey interview. *Sociological methodology* 46:1-38.

Goldstein, H., W. Browne, and J. Rasbash. 2002. Partitioning variation in multilevel models. *Understanding statistics: Statistical issues in psychology, education, and the social sciences* 1:223-231.

Graesser, A. C., Z. Cai, M. M. Louwerse, and F. Daniel. 2006. Question Understanding AID (QUAID): A web facility that tests question comprehensibility. *Public opinion quarterly* 70:3-22.Krosnick, J.A. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5:213-236.

Lenzner, T. 2014. Are readability formulas valid tools for assessing survey question difficulty? *Sociological methods & research* 43:677-698.

Olson, K., and J. D. Smyth. 2015. The effect of CATI questions, respondents, and interviewers on RT. *Journal of survey statistics and methodology* 3:361-396.

Olson, K., J. D. Smyth, and B. Cochran. 2018. Item location, the interviewer–respondent interaction, and responses to battery questions in telephone surveys. *Sociological methodology* 48:225-268.

Olson, K., J. D. Smyth, and A. Kirchner. 2019. The effect of question characteristics on question reading behaviors in telephone surveys. Forthcoming in *Journal of survey statistics and methodology*.

Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and longitudinal modeling using stata: continuous responses, 3rd Edition, Vol. 1*. College Station, TX: Stata Press.

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

Schaeffer, N.C. 2000. Asking questions about threatening topics: A selective overview. In *The science of self-report: implications for research and practice*, ed. A. A. Stone, J. B. Jobe, H. S. Kurtzman, V. S. Cain. Mahwah, 105-122. NJ: Lawrence Erlbaum Associates.

Schaeffer, N. C., and H. W. Charng. 1991. Two experiments in simplifying response categories: Intensity ratings and behavioral frequencies. *Sociological perspectives* 34:165-182.

Schaeffer, N. C., and J. Dykema. 2015. Question wording and response categories. In *International encyclopedia of the social and behavioral sciences 2nd Edition, Vol. 23*, ed. J. D. Wright, 764-770. Oxford: Elsevier.

Tourangeau, R., and T. Yan. 2007. Sensitive questions in surveys. *Psychological bulletin,* 133(5):859-883.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The psychology of survey response*. Cambridge: Cambridge University Press.

Willis, G. B. 2005. *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.

Willis, G. B., and J. T. Lessler. 1999. *Question Appraisal System: QAS-99*. Retrieved from Rockville, MD.

Yan, T. and R. Tourangeau. 2008. Fast times and easy questions: The effects of age, experience and question complexity on web survey RTs. *Applied cognitive psychology* 22:51-56.