18

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

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

CONTENTS

18.1 Introduction

Response time (RT) – the time elapsing from the beginning of question reading for a given question until the start of the next question – is a potentially important indicator of data quality that can be reliably measured for all questions in a computer-administered survey using a latent timer (i.e., triggered automatically by moving on to the next question).* In interviewer-administered surveys, RTs index data quality by capturing the entire length of time spent on a question–answer sequence, including interviewer question-asking behaviors and respondent question-answering behaviors. Consequently, longer RTs may indicate longer processing or interaction on the part of the interviewer, respondent, or both.

RTs are an indirect measure of data quality; they do not directly measure reliability or validity, and we do not directly observe what factors lengthen the administration time. In addition, either too long or too short RTs could signal a problem (Ehlen, Schober, and Conrad 2007). However, studies that link components of RTs (interviewers' question

* RTs are distinct from response latencies (RLs). RLs measure time from the end of question reading to the respondent's answer. RLs have been shown to be associated with, for example, response accuracy (Draisma and Dijkstra 2004) and task difficulty (Garbarski, Schaeffer, and Dykema 2011).

reading and response latencies) to interviewer and respondent behaviors that index data quality strengthen the claim that RTs indicate data quality (Bergmann and Bristle 2019; Draisma and Dijkstra 2004; Olson, Smyth, and Kirchner 2019). In general, researchers tend to consider longer RTs as signaling processing problems for the interviewer, respondent, or both (Couper and Kreuter 2013; Olson and Smyth 2015; Yan and Olson 2013; Yan and Tourangeau 2008).

Previous work demonstrates that RTs are associated with various characteristics of interviewers (where applicable), questions, and respondents in web, telephone, and face-to-face interviews (e.g., Couper and Kreuter 2013; Olson and Smyth 2015; Yan and Tourangeau 2008). We replicate and extend this research by examining how RTs are associated with various question characteristics and several established tools for evaluating questions. We also examine whether increased interviewer experience in the study shortens RTs for questions with characteristics that impact the complexity of the interviewer's task (i.e., interviewer instructions and parenthetical phrases). We examine these relationships in the context of a sample of racially diverse respondents who answered questions about participation in medical research and their health.

18.1.1 Response Times and Question Characteristics

Questions vary in many ways, including their structural features (e.g., number of words or clauses), difficulty (e.g., readability level), response format (e.g., yes/no, ordinal rating scale, open response), topic, and content (Dykema, et al. 2019). RTs have been shown to be related to several question characteristics, including question type (e.g., events and behaviors vs. evaluations), question length, response format, inclusion of instructions, presence of ambiguous terms, and use of fully vs. partially labeled response categories (e.g., Couper and Kreuter 2013; Olson and Smyth 2015; Yan and Tourangeau 2008). Studies of RTs and question characteristics are largely based on observational approaches (see review in Dykema, et al. 2019) in which researchers make use of a survey conducted for another purpose, code specific characteristics of the questions in the survey, and examine the association of those characteristics with RTs. The characteristics examined vary across studies as a function of the types of questions available in the questionnaire and researcher interests. Replication across surveys, topics, and populations is critically important, given that many question characteristics are study-specific and collinear (Schaeffer and Dykema forthcoming).

In this chapter, we examine the association between RTs and question characteristics available in our own observational study. Table 18.1 provides the list of question characteristics and hypotheses. We base our hypotheses on relationships demonstrated in previous research and expectations about whether the characteristic is likely to increase the cognitive processing burden of the respondent, interviewer, or both. Some hypotheses are evident; others require explication. See Online Appendix 18A for background and justification regarding H1a–H1l. We formulate hypotheses under the assumption that other question characteristics are held constant.

In addition to the individual or "ad hoc" question characteristics described above, we also examine the association of several established question evaluation tools with RT, including the Flesch–Kincaid grade level, the Question Understanding Aid (QUAID; Graesser, et al. 2006), the Question Appraisal System (QAS; Willis 2005; Willis and Lessler 1999), and the Survey Quality Predictor (SQP; Saris and Gallhofer 2007) (see Online Appendix 18B). Each tool identifies multiple question characteristics that may be problematic for

TABLE 18.1

Hypotheses about the Effect of Question Characteristics on Response Times

Hypothesis	Question Characteristic	Effect on RTs
H1a	Number of words	+
H1b	Question order	–
H1c	Question type	Demographics < events/behaviors < subjective
H1d	Question form	Yes/no < unipolar ordinal, bipolar ordinal, nominal, discrete value
H1e	Definition in the question	+
H1f	List-item question	+
H1g	Sensitive question	–
H1h	Race-related question	+
H1i	Battery structure	First in battery > later; First in series > later
H1j	Emphasis in the question	–
H1k	Interviewer instructions	+
H1l	Parenthetical phrases	–
H2a, 3a	Flesch–Kincaid grade level	+
H2b, 3b	QUAID problem score	+
H2c, 3c	QAS problem score	+
H2d, 3d	SQP quality score	–
H4a	Interaction of number of interviews by interviewer instructions	–
H4b	Interaction of number of interviews by parenthetical phrases	–

Notes: H1a–H1l and H3a–H3d are net of the effects of other question characteristics; H2a–H2d are for bivariate relationships.

respondents or interviewers, and the tools can be used to code questions and characteristics from any type of survey. Although the tools differ in their implementation and scope, they can be used to produce a question-level "problem" or "quality" score that indicates the complexity of the question. We expect that more complex questions (as indicated by scores from the established tools) are associated with longer RTs because they are harder for interviewers to read and harder for respondents to answer (Table 18.1 H2a to H2d; H2d is negative because a higher SQP quality score indicates less complexity). Consistent with expectations, Olson and Smyth (2015) reported that questions with higher reading levels (harder to read) took longer to administer. We are not aware of studies that examine the relationship between the other tools and RTs. (Yan and Tourangeau [2008] examined the relationship between individual question characteristics and QUAID, but they did not include QUAID as a predictor of RT.)

Coding individual question characteristics and generating scores using the established tools is time-consuming and can be costly. Thus, whether the individual characteristics and scores from established tools each independently account for variance in RTs or are duplicative of each other is of interest. We evaluate this by examining whether scores from the established tools predict RTs net of individual question characteristics (H3a to H3d in Table 18.1): although some aspects of the characteristics that are coded to produce these scores overlap with individual question characteristics (e.g., question length), they

also incorporate features beyond the individual characteristics with potential implications for RTs.

18.1.2 Response Times and Interviewers' Experience

An important interviewer characteristic to consider in predicting RTs is the interviewer's level of experience. Interviewers appear to increase their pace within an interview (as they gain experience with an individual respondent), within a study (as they gain experience with the particular questionnaire), and across studies (as they become more experienced in general). Their faster speed may be because they develop shortcuts (e.g., alter questions or decrease standardized practices), become more fluent, head-off problems, and so forth (Bergmann and Bristle 2019; Böhme and Stöhr 2014; Holbrook, et al. Chapter 17; Kirchner and Olson 2017; Olson and Peytchev 2007; Olson and Smyth Chapter 20).

In this chapter, we are primarily concerned with within-study experience (i.e., the number of interviews interviewers have conducted). Previous research indicates that the time to complete an entire interview (the aggregate of RTs) and interviewer reading times decrease with the number of interviews completed for a given study (Bergmann and Bristle 2019; Kirchner and Olson 2017; Loosveldt and Beullens 2013; Olson and Peytchev 2007), particularly for inexperienced interviewers (Olson and Peytchev 2007), and accounting for changes in the types of respondents interviewers encounter over the course of the field period (Kirchner and Olson 2017).

We propose that interviewer experience interacts with question characteristics that primarily impact interviewers' task complexity (Olson and Smyth 2015) in predicting RTs because these are the characteristics for which interviewers have the most discretion. In this study, these question characteristics include interviewer instructions and parenthetical phrases. As they complete interviews and become more familiar and comfortable with the questionnaire, we expect interviewers will decrease their attention to and reading of interviewer instructions and be less likely to incorporate discretionary parenthetical phrases. Thus, with increasing interviewer experience (more interviews completed), RTs will decline more rapidly for questions with instructions or parenthetical phrases than without them (H4a and 4b; Table 18.1).

18.2 Data and Methods

Data for this study are from the Voices Heard computer-assisted telephone interview (CATI) survey, which was designed to measure perceptions of barriers and facilitators to participating in medical research studies that collect biomarkers (e.g., saliva and blood) among respondents from various racial and ethnic groups (White, Black, Latino, and American Indian). We employed a quota sampling strategy because screening to identify members in non-White groups would have been prohibitively expensive. The quota sample consisted primarily of volunteers but also used a targeted list of names provided by a commercial vendor (see Online Appendix 18C for more detail). Interviewers conducted 410 usable interviews (in English only) with an average length of 25.21 minutes between October 2013 and March 2014. Respondents received a $20 cash incentive. The 96 questions included in the survey asked about: likelihood to participate in medical research based on the type of study (e.g., to collect tissue) and characteristics of requestor (e.g., "a

member of your community"); things medical researchers do to encourage participation (e.g., provide results); concerns about participating in medical research; attitudes toward medical researchers; health status, health-related quality of life, health behaviors and conditions, and health care use; knowledge of research procedures; and social and demographic characteristics.

18.2.1 Measures

Dependent variable. RTs were collected by the CATI computer software as the amount of time (in seconds) spent on each question (mean 13.22 seconds, standard deviation 8.96, range 1–110). Values were top- and bottom-coded at the 99th and 1st percentiles within each item and log-transformed to correct for outliers and skew (Yan and Tourangeau 2008).

Individual question characteristics. Research assistants coded the previously identified individual question characteristics (H1a–H1l in Table 18.1) under the direction of the authors; no interrater reliability statistics were calculated, but codes were verified by the first author. Descriptive statistics for question characteristics are provided in the first column of Table 18.2.

Established tools for evaluating questions. We measured readability using the Flesch–Kincaid grade level. A higher level indicates the question's text is more difficult to read. For QUAID, we tallied the number of problems flagged by the online tool across five comprehension difficulty categories. QAS was coded by a member of the research team and operationalized as a composite sum of the number of problems identified out of 27 possible problems. SQP was coded by an undergraduate research assistant using SQP's online documentation. We use SQP's "quality estimate" (the product of a question's estimated reliability and validity) (see Online Appendix 18B).

Interviewer and respondent characteristics. The key interviewer characteristic of interest is within-study experience (number of interviews the interviewer completed up to the current interview). Other interviewer characteristics included as controls are: race (White, non-White [very few interviewers were Black, Latino, or Asian]), gender, age, and prior interviewing experience (less than one year or one year or more). Respondent characteristics included as controls are: race/ethnicity (Black, Latino, American Indian, and White), gender, age, and education (high school education or less, some college, and college or more). The last two characteristics are used in prior studies to examine or control for factors associated with response processing and cognitive ability (see Online Appendix 18D, Table A18.D1).

18.2.2 Analytic Strategy

The analytic sample includes 410 respondents asked 95 or 96* questions by one of 24 interviewers, yielding 39,052 question–answer sequences, which are the unit of analysis. We use cross-classified random-effects linear regression models to predict the log-transformed RTs using Stata 15.1. We use the *mixed* command with restricted maximum likelihood (*reml*) to analyze the data with a variance structure that uses crossed random effects to account for the fact that RT for each question is measured for each respondent and interviewer, and

* One question was a follow-up to a filter question that was not asked if respondents answered "yes" to the filter question.

TABLE 18.2

Descriptive Statistics and Regression Results of Response Times on Characteristics of Questions, Interviewers, and Respondents, Voices Heard Study

	Descriptive Statistics				Regression		
Question Characteristics	**Mean or Percent**	**Std. Dev.**	**Min.**	**Max.**	**Coef.**	**Std. Err.**	
Number of words	30.47	16.17	5.00	75.00	0.018	0.003	***
Question order	48.50	27.86	1.00	96.00	−0.002	0.002	
Question type							
Event or behavior (reference category)	57.3%						
Subjective	28.1%				0.211	0.147	
Demographic	14.6%				0.231	0.155	
Question form							
Yes/no (reference category)	30.2%						
Nominal	8.3%				0.208	0.123	
Discrete value	2.1%				0.328	0.188	
Bipolar ordinal	16.7%				1.067	0.170	***
Unipolar ordinal	42.7%				0.600	0.138	***
Definition in the question (vs. not)	5.2%				0.079	0.159	
List-item question (vs. not)	35.4%				0.027	0.063	
Sensitive question (vs. not)	10.4%				0.073	0.088	
Race-related question (vs. not)	9.4%				−0.017	0.100	
Battery structure							
First in battery	9.4%				0.102	0.114	
Later in battery (reference category)	44.8%						
First in series	6.3%				0.275	0.127	*
Later in series	31.3%				0.165	0.105	
Stand-alone	8.3%				0.242	0.139	
Emphasis in the question (vs. not)	19.8%				−0.316	0.102	**
Interviewer instructions (vs. not)	9.4%				0.201	0.112	
Parenthetical phrases (vs. not)	34.4%				−0.336	0.082	***
Flesch–Kincaid grade level	12.22	5.16	0.00	22.10	0.018	0.008	*
QUAID problem score	4.38	2.30	1.00	12.00	0.012	0.014	
QAS problem score	1.00	1.02	0.00	4.00	−0.033	0.045	
SQP quality score	0.50	0.05	0.44	0.67	−0.319	0.720	
Intercept					1.166	0.443	**
Random-effects parameters							
Interviewer-level variance					0.003	0.001	*
Question-level variance					0.045	0.007	***
Respondent-level variance					0.012	0.001	***
Residual variance					0.085	0.001	***
Wald chi-square					693.83	(df 35)	***
Log-restricted likelihood					−8,268.60		

Notes: Std. Dev. = standard deviation, Min. = minimum, Max. = maximum, Coef. = coefficient, Std. Err. = standard error. Descriptive statistics are calculated at the level of the question (N = 96) for question characteristics. Regression analysis is conducted at the level of the question–answer sequence (N = 39,052). Regression model also controls for respondent (race/ethnicity, gender, age, and education; N = 410) and interviewer characteristics (race/ethnicity, gender, age, prior interviewing experience, and study-specific experience; N = 24).

*p < 0.05, **p < 0.01, ***p < 0.001.

respondents are nested within interviewers. The base model predicting RT i for question j_1, respondent j_2, and interviewer k is $\ln(\text{Response time})_{i(j_1,j_2)k} = \beta_0 + u_{j_1} + u_{j_2} + v_k + e_{i(j_1,j_2)k}$. In this model, $u_{j_1} \sim N(0,\sigma^2_{u(1)})$, $u_{j_2} \sim N(0,\sigma^2_{u(2)})$, $v_k \sim N(0,\sigma^2_v)$, and $e_{i(j_1,j_2)k} \sim N(0,\sigma^2_e)$).

The full model predicting RT includes a series of fixed effects for questions, respondents, and interviewers:

$$\ln(\text{Response time})_{i(j_1,j_2)k} = \beta_0 + \sum_{b=1}^{B}\beta_b \text{Question characteristics}_{j_1k}$$

$$+\sum_{c=1}^{C}\beta_c \text{Respondent characteristics}_{j_2k} + \sum_{d=1}^{D}\beta_d \text{Interviewer characteristics}_k$$

$$+u_{j_1} + u_{j_2} + v_k + e_{i(j_1,j_2)k}$$

Because RTs are (natural) log-transformed, the coefficients can be interpreted in terms of percentage change, such that RTs change by 100 * [exp(β) − 1] percent for a one-unit increase in the independent variable, holding all other variables in the model constant.

18.3 Results

Table 18.2 presents a full model that regresses RTs on characteristics of questions, interviewers, and respondents (see Online Appendix 18D, Table A18.D2 for results from the partial models). Several of the significant effects of individual question characteristics align with our expectations (Table 18.1), net of the other characteristics. Each additional word in the question is associated with a 1.8% increase (i.e., 100 * [exp(.018) − 1]) in RTs (Table 18.2), consistent with H1a. Increasing question order is associated with a decrease in RT when the model does not control for scores from established tools for evaluating questions (Online Appendix 18D, Table A18.D2, Model 1), but this effect is not significant in the full model (Table 18.2), so H1b is not supported in the full model. Questions that have bipolar or unipolar ordered categories have longer RTs than yes/no questions (the reference group), but nominal and discrete-value questions are not significantly different from yes/no questions, so H1d is partially supported.* We find no evidence supporting H1c (question type), H1e (definition), H1f (list item), H1g (sensitive), H1h (race-related), and H1i (battery structure). However, when the question includes emphasis (i.e., bolded text), RTs are shorter, consistent with the expectation that emphasis aids in respondents' processing efficiency (H1j).

The hypotheses focused on question characteristics that impact the complexity of the interviewer's task are partially supported. The presence of an interviewer instruction is associated with increased RTs in the model that examines individual question characteristics

* The discrete-value questions ask respondents to report numerical answers: year of birth and number of days they drank alcohol in the past month. RTs are lower for all question forms compared to bipolar ordinal questions ($p<.001$) and lower for nominal questions compared to unipolar ordinal questions ($p<.01$) (not shown).

(Online Appendix 18D, Table A18.D2, Model 1), but it is not significant when controlling for scores from established tools for evaluating questions (Table 18.2), so H1k is not supported in the full model. The presence of a parenthetical phrase is also associated with decreased RTs, consistent with H1l and the expectation that, on average, interviewers read parenthetical phrases only when deemed necessary rather than with every question administration (Olson, Smyth, and Kirchner 2019).

When we examine the association between RTs and scores from the established tools to evaluate survey questions – important because investigators might only use one measure – we find that Flesch–Kincaid grade level (H2a) and QUAID problem score (H2b) are each positively associated with RTs (Online Appendix 18D, Table A18.D2, Models 2 and 3). Thus, hypotheses concerning bivariate relationships are supported for Flesch–Kincaid grade level and QUAID problem score, but not for QAS problem score or SQP quality score. When the individual question characteristics are included in the model with the established tools (Table 18.2), the effect of QUAID problem score is attenuated and not statistically significant, while the effect of grade level is attenuated but still significant (each additional grade level is associated with a 100 * [exp(.018) − 1] = 1.8% increase in RTs). Thus, H3 is only supported for the Flesch–Kincaid grade reading level (H3a).

The full model reduces the question-level variance relative to the base model by 87%, while the model with the individual question characteristics (that is, without the established tools) reduces the question-level variance by 86% (Online Appendix 18D, Table A18.D3). Thus, most of the question-level variation in RTs in these data is explained by this set of individual question characteristics. In the model that controls for the characteristics of respondents and interviewers, RTs are significantly different and longer for women, older respondents, and Latino respondents compared to other racial and ethnic groups (Online Appendix 18D, Table A18.D1); these effects remain largely unchanged in the models that also control for the question characteristics and established tools for evaluating questions (not shown).

Next we turn to our hypotheses that RTs will decrease faster with an increasing number of interviews completed in questions with interviewer instructions or parenthetical phrases than in questions without these characteristics. As predicted, there is a significant negative interaction of number of interviews completed with interviewer instructions and parenthetical phrases (results available upon request). Figures 18.1 and 18.2 show the predicted marginal means of log-transformed RTs with increasing numbers of interviews completed for questions with and without interviewer instructions and parenthetical phrases. Questions with interviewer instructions have longer RTs than those without interviewer instructions, but, as interviewers conduct more interviews, RTs decrease more rapidly for questions with instructions than without (Figure 18.1). The interaction effect is significant ($p<.05$) overall, yet its significance varies across the span of number of interviews completed: the marginal effect of having interviewer instructions as part of the question (vs. not) is statistically significant ($p = .046$) with the first interview and drops below statistical significance ($p = .052$) by the fourth interview (not shown). Thus, H4a is supported – for the first few interviews. Figure 18.2 shows that RTs are shorter, on average, when questions include parenthetical phrases compared to when they do not, and although RTs decrease for questions with and without parenthetical phrases over the number of interviews completed, the slope is steeper for questions that contain parenthetical phrases. The marginal effect is significant across the span of number of interviews completed (not shown). Thus, H4b is supported.

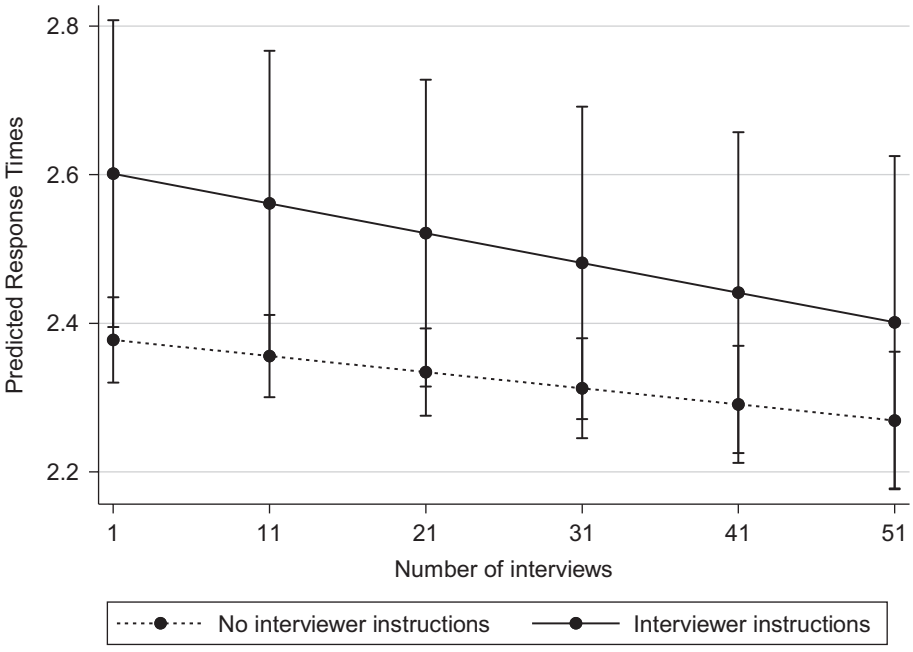

FIGURE 18.1
Predicted marginal means of (log-transformed) response times by the number of interviews completed and interviewer instructions in the question.

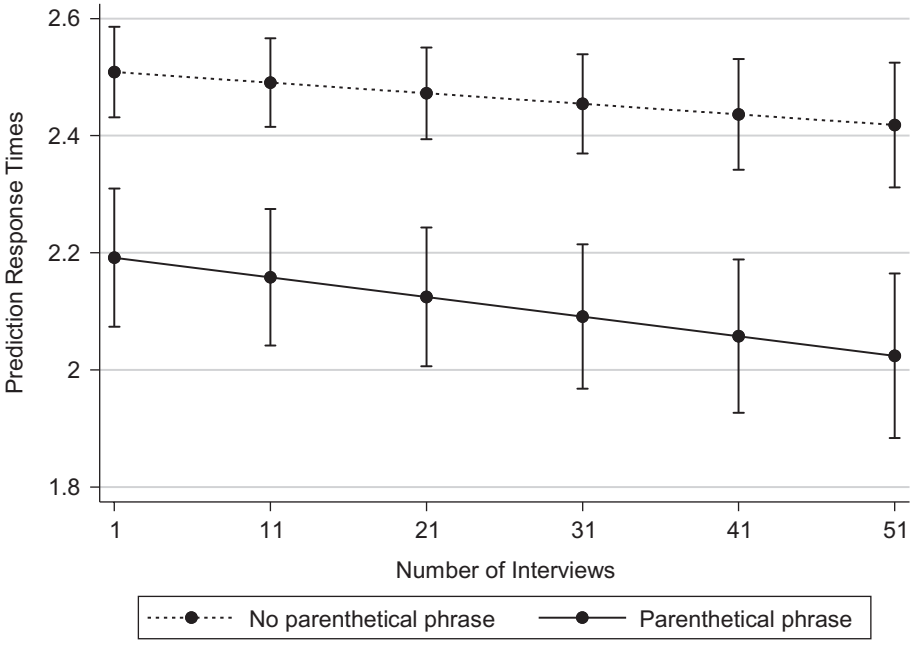

FIGURE 18.2
Predicted marginal means of (log-transformed) response times by the number of interviews completed and parenthetical phrases in the question.

18.4 Discussion

This study examines how RTs are associated with characteristics of questions, interviewers, and respondents in a sample of racially diverse respondents answering questions about participating in medical research and health. Results add to the findings about question characteristics that are associated with RTs, our indicator of data quality. Results show that some individual question characteristics are associated with RTs in expected ways (word count, ordinal vs. yes/no question forms, emphasis in questions, parenthetical phrases), as were some established tools (i.e., Flesch–Kincaid grade level and QUAID problem score). However, only grade-level readability remained significant in the full model that also controlled for the individual question characteristics. Although its utility as a tool for measuring question complexity is disputed (Lenzner 2014), the Flesch–Kincaid grade level score independently predicts RTs in this and other studies (Olson and Smyth 2015) – future research should focus on identifying the mechanism through which this occurs (e.g., increased interaction) and whether grade-level readability predicts other data quality measures. Overall, contrary to our expectations, the other evaluation tools did not capture additional complexity in questions that predicted RTs beyond the individual question characteristics. While QUAID, QAS, and SQP are useful for improving questions prior to data collection and are associated with several data quality measures other than RTs (Dykema, et al. 2019; Forsyth, Rothgeb, and Willis 2004; Maitland and Presser 2016; Olson 2010; Olson, Smyth, and Kirchner 2019; van der Zouwen and Smit 2004), they did not contribute to explaining variation in RT in this study. The methods researchers use to operationalize scores from these tools vary across studies. Future research should examine the implications of different operationalizations.

Questions with emphasis (e.g., bolding of text) were associated with reduced RTs, consistent with the notion that emphasis aids in respondents' cognitive processing. We note that emphasis might also increase cognitive processing demands and thus question reading time for interviewers as hypothesized by Olson and Smyth (2015), but our study suggests that the net effect on RTs is a decrease. The example of emphasis indicates that the component parts of RTs (i.e., interviewer and respondent contributions) should be examined when hypotheses about the mechanisms producing the effects of question characteristics conflict across actors or interactional sequences (e.g., shorter for respondents but longer for interviewers, or producing interactional moments that lead to shorter or longer responses as a result). Although question-level RTs are a useful and easily accessible measure of data quality, truly understanding certain question characteristics will require a more nuanced – and more labor-intensive – analysis that decomposes the component parts of RT by actor and possibly even type of behavior.

In this study, some of the effects of emphasis on RTs might be driven by the dependency of emphasis on questions that are structurally interrelated because they are part of a battery; that is, the questions with emphasis were in batteries and the emphasis was likely needed to distinguish among the items in the batteries. The intersection of batteries and emphasis in questions illustrates an important issue with respect to the observational study of question characteristics, however – question characteristics are not independent of each other. This has implications for both the meaning of a question characteristic and its association with data quality. As Dykema and colleagues (2019) point out, many studies of this kind have not taken structural dependencies into account in analysis, at least not systematically; that is the case in this study as well. In observational studies in particular, the joint distribution of question characteristics affects whether group sizes are sufficient to estimate main effects and interactions (Dykema, et al. 2016). The results of these types of

studies may depend on which combinations of characteristics were accounted for in each particular analysis, which may contribute to the lack of replicability in the effects of certain question characteristics (individual characteristics or established tools that combine multiple characteristics into more comprehensive scores) across studies, complicating findings of which question characteristics are better predictors of data quality. Future research should include more study replications under different survey conditions and experimental designs to parse the dependencies where possible, especially for those characteristics for which existing findings are the least consistent.

Overall, interviewers' experience within the survey (i.e., number of interviews completed) is not associated with RTs. This may be due to the telephone mode, which is more monitored compared to face-to-face modes (Kirchner and Olson 2017). However, we found that interviewers' experience in the study interacted with key question characteristics – interviewer instructions and parenthetical phrases – that are used at or attended to with the interviewer's discretion. Specifically, questions with interviewer instructions have longer RTs for interviewers with fewer interviews and the slope decreases more rapidly for questions with interviewer instructions compared to questions without. However, the difference in slope is only significant for the first three interviews, indicating that interviewers no longer read or attend to these instructions after the first few interviews and may apply them from memory. In contrast, the inclusion of parenthetical phrases in questions served to decrease RTs at a significantly steeper rate with more interviews completed. This relationship may indicate that interviewers are treating parenthetical phrases as optional within interviews (as the unconditional effect of parentheticals indicated), and increasingly so as they complete more interviews during the study, with the result of lowering average RTs over time (that is, with more interviews completed).

If interviewers treat the parenthetical phrases as optional during question reading, this has implications for standardized survey administration, as all respondents are not hearing the same question. It raises the question of whether the parenthetical or non-parenthetical version of the item counts as the scripted administration (e.g., Olson, Smyth, and Kirchner 2019). However, as interviewers complete more interviews, they may be more adept at learning and facilitating small micro-adjustments to question asking, such as whether and when to omit parenthetical phrases or attend to instructions, that are aligned with the goals of standardization and keeping with conversational practices and maintaining rapport – if not necessarily aligned with the rules of standardization (Garbarski, Schaeffer, and Dykema 2016). This speaks to the notion of interviewers as pragmatists who work to complete the interview and learn the complexity of the interview task over time (Paul Beatty 2019, personal communication). As we do for respondents, future research must consider visual design (Dillman, Smyth, and Christian 2014) as integral for interviewers in terms of whether and how they read and attend to parts of the instrument; at least with interviewers involved, we can train on attention to various cues and retrain if standardized practices diminish over time. With respect to parenthetical phrases in particular, the evidence here and elsewhere is becoming clear: they are associated with indicators of lower data quality (Dykema, et al. 2016, 2019; Olson, et al. 2019).

With regard to limitations of the study, we note that RT is an indirect measure of data quality: we can presume that interviewers are choosing to not read parenthetical phrases, and increasingly so as they complete more interviews, but we are not directly observing behavior in this study. The strength of a measure of RT is that it is low-cost and easily obtained for every question–answer sequence in the data (with the correct programming capabilities to capture it). In terms of its validity, however, there is more work to be done to examine what actually underlies RT as a measure of data quality. For example, behavior coding could be used to examine what behaviors are associated with longer or shorter RTs.

Such studies would lend more credibility to using RTs as an indirect measure of data quality in more studies. Indeed, the results of any study depend on the quality of the criteria at hand, both the dependent and independent variables.

An additional limitation is that respondents were not recruited randomly due to cost and feasibility constraints, which limits the generalizability of our sample to a larger population of respondents. As with other observational studies of question characteristics, this issue exists at the level of the question (characteristics are not randomly sampled from the universe of all question characteristics but rather fit for the purpose of a given study) as well as at the level of the interviewer (who are employees at one particular survey organization, and not randomly assigned to cases but rather assigned due to proximity, shift, and so forth). Another factor that would be useful to know is respondents' status as English language speakers (e.g., is it their primary language spoken), for which we do not have information.

This research advances the field of survey methodology by examining RTs in the context of different question characteristics and established tools for evaluating questions as well as specific interviewer characteristics within a uniquely diverse sample of respondents. The results have direct implications for survey measurement, questionnaire design, interviewing methods, and interviewer training. The results expand our understanding of the joint influence of characteristics of questions, interviewers, and respondents – the first two of which may be modifiable in the course of survey research – and their application to the development of practical methods for improving the quality of survey data.

Acknowledgments

The collection of survey data for the Voices Heard Survey was funded by NIMHD grant P60MD003428 (PD: A. Adams). Project: Increasing Participation of Underrepresented Minorities in Biomarker Research (PI: D. Farrar Edwards). This study is based upon work supported by the National Science Foundation (grant number SES-1853094 to J. Dykema and D. Garbarski]. Project: Effects of Interviewers, Respondents, and Questions on Survey Measurement. Additional support was provided by the University of Wisconsin Survey Center (UWSC), which receives support from the College of Letters and Science at the University of Wisconsin-Madison; the facilities of the Social Science Computing Cooperative and the Center for Demography and Ecology (NICHD core grant P2C HD047873) at the University of Wisconsin-Madison; and the Graduate School and the Gannon Center for Women and Leadership at Loyola University Chicago. The authors thank the editors of this volume for their helpful comments on earlier drafts. Opinions expressed here are those of the authors and do not necessarily reflect those of the sponsors or related organizations.

References

Bergmann, M., and J. Bristle. 2019. Reading fast, reading slow: The effect of interviewers' speed in reading introductory texts on response behavior. *Journal of Survey Statistics and Methodology*. Advanced Access.

Böhme, M., and T. Stöhr. 2014. Household interview duration analysis in CAPI survey management. *Field Methods* 26(4):390–405.

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(1):271–286.

Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 4th edition. Hoboken, NJ: Wiley & Sons Inc.

Draisma, S., and W. Dijkstra. 2004. Response latency and (para)linguistic expression as indicators of response error. In: *Methods for Testing and Evaluating Survey Questionnaires*, ed. S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 131–148. New York: Springer-Verlag.

Dykema, J., N. C. Schaeffer, D. Garbarski, and M. Hout. 2019. The role of question characteristics in designing and evaluating survey questions. In: *Advances in Questionnaire Design, Development, Evaluation, and Testing*, ed. P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot, 119–152. Hoboken, NJ: Wiley.

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.

Ehlen, P., M. F. Schober, and F. G. Conrad. 2007. Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Processes* 44(3):245–265.

Forsyth, B., J. M. Rothgeb, and G. B. Willis. 2004. Does pretesting make a difference? An experimental test. In: *Methods for Testing and Evaluating Survey Questionnaires*, ed. S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 525–546. New York: Springer-Verlag.

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.

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):1–38.

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(1):3–22.

Kirchner, A., and K. Olson. 2017. Examining changes of interview length over the course of the field period. *Journal of Survey Statistics and Methodology* 5:84–108.

Lenzner, T. 2014. Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods and Research* 43(4):677–698.

Loosveldt, G., and K. Beullens. 2013. The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research* 42(6):1422–1430.

Maitland, A., and S. Presser. 2016. How accurately do different evaluation methods predict the reliability of survey questions? *Journal of Survey Statistics and Methodology* 4(3):362–381.

Olson, K. 2010. An examination of questionnaire evaluation by expert reviewers. *Field Methods* 22(4):295–318.

Olson, K., and A. Peytchev. 2007. Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly* 71(2):273–286.

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

Olson, K., J. D. Smyth, and A. Kirchner. 2019. The effect of question characteristics on question reading behaviors in telephone surveys. *Journal of Survey Statistics and Methodology*. Advanced Access.

Saris, W. E., and I. N. Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley.

Schaeffer, N. C., and J. Dykema. Forthcoming. Advances in the science of asking questions. *Annual Review of Sociology*.

van der Zouwen, J., and J. H. Smit. 2004. Evaluating survey questions by analyzing patterns of behavior codes and question-answer sequences: A diagnostic approach. In: *Methods for Testing and Evaluating Survey Questionnaires*, ed. S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 109–130. New York: Wiley.

Willis, G. B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

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

Yan, T., and K. Olson. 2013. Analyzing paradata to investigate measurement error. In: *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. F. Kreuter, 73–96. Hoboken, NJ: John Wiley & Sons.

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(1):51–68.