2015

The Effect of CATI Questions, Respondents, and Interviewers on Response Time

Kristen Olson
University of Nebraska-Lincoln, kolson5@unl.edu

Jolene D. Smyth
University of Nebraska-Lincoln, jsmyth2@unl.edu

Published in *Journal of Survey Statistics and Methodology* 3 (2015), pp. 1–36
doi: 10.1093/jssam/smv021

The Effect of CATI Questions, Respondents, and Interviewers on Response Time

Kristen Olson and Jolene D. Smyth

Department of Sociology, University of Nebraska–Lincoln

Corresponding author – Kristen Olson, Department of Sociology, University of Nebraska– Lincoln, Lincoln, Nebraska, USA; email kolson5@unl.edu

Abstract
In this paper, we evaluate the joint effects of question, respondent, and interviewer characteristics on response time in a telephone survey. We include question features traditionally examined, such as the length of the question and format of response options, and features that have yet to be examined that are related to the layout and format of intervieweradministered questions. We examine how these question features affect the time to ask and answer survey questions and how different interviewers vary in their administration of these questions. This paper uses paradata from the Work and Leisure Today survey and uses cross-classified random effects models. Overall, most of the variation in response time is due to question characteristics, rather than respondent or interviewer attributes. Additionally, we find that question characteristics related to necessary survey design features and respondent confusion are the primary predictors of response time, with little effect of visual design features of the question.We also find modest differences in the effects of question characteristics by interviewer experience.

Keywords: CATI surveys, Questionnaire design, Response latency

This work was supported by the National Science Foundation [SES-1132015]. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. An earlier version of this paper was presented at the Annual Meeting of the American Association for Public Opinion Research in Anaheim, California, in May 2014. The authors thank the editor and two anonymous reviewers from JSSM for helpful comments that improved the manuscript.

1. Introduction

Designing questionnaires requires survey researchers to make many decisions about question content, wording, length, ordering, grouping, and other characteristics (Bradburn, Sudman, and Wansink 2004; Dillman, Smyth, and Christian 2014). Experimental evaluations of these design features generally focus on one characteristic at a time, often question wording (Schuman and Presser 1981; Fowler and Mangione 1990; Fowler 1995). This limits the number of features that feasibly can be evaluated and discourages evaluation of features that are common but difficult to experimentally manipulate (e.g., question reading level). In addition, by design, experiments impede our understanding of the joint effects of multiple design features (Dillman 1991), and whether the effects of these design features uniformly operate for all interviewers or respondents. This paper jointly evaluates the effect of multiple question features in a CATI survey, and how these design features vary in effects for experienced versus inexperienced interviewers.

Two recent studies have used multilevel analytic techniques to examine the joint effects of multiple question features in web (Yan and Tourangeau 2008) and in-person surveys (Couper and Kreuter 2013). Importantly, these studies incorporated respondent (and, where appropriate, interviewer) characteristics into their analyses because respondents and interviewers bring their own knowledge, experience, and abilities to bear on the question/answer process. No known studies have applied similar methods to understanding the design of telephone questionnaires.

Evaluating the joint effects of question, respondent, and interviewer characteristics across items in a questionnaire requires a data quality metric that is shared across items. Low item nonresponse rates in telephone surveys make it a difficult outcome to use for this purpose. Another possibility is response time, which has been previously used as a proxy measure of data quality, and is of interest for factors that contribute to the overall length of a questionnaire (Olson and Parkhurst 2013; Yan and Olson 2013).

In this paper, we examine the joint effects of question, respondent, and interviewer characteristics on response time in a computer-assisted telephone interview (CATI) questionnaire. Following the work of Yan and Tourangeau (2008) and Couper and Kreuter (2013), we use multilevel models that account for the joint effects of questions, respondents, and interviewers. We expand on existing work in four important ways. First, we use a telephone survey. Others have examined response time in telephone surveys (e.g., Bassili and Fletcher 1991; Bassili 1996; Mulligan, Grant, Mockabee, and Monson 2003; Johnson 2004), but little attention has been given to the joint effects of interviewer, respondent, and question characteristics in this mode. Telephone surveys are meaningfully different from web surveys in that there is an interviewer, which is thus an influence not examined by Yan and Tourangeau (2008). Couper and Kreuter (2013) examine the effects of interviewers, respondents,

and questions in a face-to-face survey with a self-administered ACASI component on at least one-fourth of the questions. The ACASI component may attenuate the effects of the interviewer relative to a mode with no ACASI component. Additionally, telephone surveys have restricted communication channel capacity compared to face-to-face interviews because the interviewer and respondent do not see each other (e.g., de Leeuw 1992, 2005). Moreover, much previous research on CATI surveys has used “active” (i.e., activated by interviewers) timers rather than “latent” timers used to collect survey paradata. Second, we apply literature on the visual design of self-administered questionnaires (e.g., Jenkins and Dillman 1997; Christian and Dillman 2004) to interviewer-administered surveys as a guide for the question characteristics included in the model. Very little research has examined the effect of visual design in CATI surveys (for an exception, see Edwards, Schneider, and Brick 2008). This means that our collection of 21 question characteristics examined is more expansive than that considered by Yan and Tourangeau (2008; 6 question characteristics) or Couper and Kreuter (2013; 8 question characteristics). Third, we use a different modeling approach than previous work in that response times are crossclassified by respondents and questions (following Yan and Tourangeau 2008), with both respondents and questions nested within interviewers (following Couper and Kreuter 2013). Fourth, we evaluate whether question characteristics have different effects on response time for experienced versus inexperienced interviewers.

In sum, we examine the following three questions:

(1) How much variability in CATI survey response time is due to interviewers, respondents, and questions?
(2) What question, respondent, and interviewer characteristics are associated with response time?
(3) Do experienced versus inexperienced interviewers vary in their effects on response time for different question characteristics?

2. Background

Response times have been used as indicators of potential problems in the response process that might be linked to measurement error (Couper 1998; Yan and Olson 2013). For example, answering too fast may indicate that respondents are not fully carrying out the response process (Malhotra 2008), and answering too slowly may indicate that they are having difficulty with the question (Bassili and Fletcher 1991; Yan and Tourangeau 2008; Couper and Kreuter 2013). Thus, examining the factors that contribute to response time can help surveyors understand the factors that may contribute to measurement error in survey responses.

2.1 Measuring Response Time

For CATI surveys, response time measures come from paradata records that capture the duration from when the question appears on the interviewer's screen to when a response is entered and the CATI program moves to the next question. This is a "latent" timer; the interviewer does not directly interact with the timing device (Mulligan et al. 2003). Most of the early work on response times used active timers that were triggered by interviewers (e.g., Bassili 1996), with a focus on the time that it takes the respondent to answer a question after the interviewer stopped reading it. Latent timer durations combine the influence of the respondent, the interviewer, and the questionnaire into one measure. In this paper, we parse out how much of the variance in response time is due to questions, respondents, and interviewers. The specific design features and respondent and interviewer characteristics that ought to affect response time are described below and summarized in Table 1.

2.2 Question Features

Many features can affect response time in a CATI questionnaire. Some are easily controlled by survey designers, but others are more difficult to control without changing the topic of a survey. Some question features will make questions more confusing, complex, or difficult to administer. Still others may impact the efficiency with which respondents can process the questions or interviewers can read the questions.

2.2.1 Necessary question features. We refer to characteristics of questions that are largely determined by the survey topic and analytic goals as necessary question features. Questionnaire designers have limited abilities to alter these features. For example, some constructs can be measured with few words (e.g., age), while others will require more words (e.g., hours of volunteer work in the past week). Likewise, a true yes/no question will have only two substantive response options, and a question asking for month of birth can have only up to 12 response options. Of course, researchers do have some discretion with these features. They can use many or few words in question stems, offer five or seven points in an ordinal scale, and choose what level of detail to use for nominal categories (e.g., religions, marital statuses, etc.). But, generally, survey researchers do not have full discretion and even these decisions should be strongly driven by measurement properties of the items and analytic needs. Other examples of necessary question features are the type of information requested and format of response options.

The first necessary question features are the number of words in a question and the number of response options. Longer questions should increase reading time and response time (Yan and Tourangeau 2008; Couper and Kreuter 2013). Likewise, questions with many response options should take longer to answer because there is more information to process (Yan and Tourangeau 2008).

Table 1. Summary of Question, Respondent, and Interviewer Characteristics

Question characteristics	
Question sequence number Length of the question (i.e., number of words) Number of response options available Question type (e.g., attitude, behavior, demographic) Response options format (e.g., open, closed, etc.)	Necessary question features (i.e., features largely dictated by measurement needs)
Question reading level Mismatch between question and response options The question contains terms that are likely unknown The question is on a sensitive topic	Features expected to affect respondent task complexity
The question contains interviewer instructions A probe is displayed on the question screen The question text contains parentheses The question is asked over two screens The question has visual emphasis (i.e., bold, italics, etc.) Interviewer backed up on the question	Features expected to affect interviewer task complexity
The question is the first in a battery The question is later in a battery A definition is provided with the question A transition statement is provided with the question The question feeds into a skip instruction The question is a follow-up item in a skip instruction	Features expected to affect respondent processing efficiency
Respondent Characteristics	
Respondent age Respondent education Respondent sex Respondent currently employed Respondent has a laptop, desktop, or tablet computer	
Interviewer Characteristics	
Interviewer sex Interviewer race Interviewing experience Workload	

Following previous research, we expect variability across different types of questions. We expect demographic questions to be answered most quickly because demographic information is commonly known and easily retrieved from memory (Bassili and Fletcher 1991; Tourangeau, Rips, and Rasinski 2000; Yan

and Tourangeau 2008). Behavioral questions ought to require more retrieval than demographic questions, but the needed information is generally fairly accessible. Consequentially, behavioral questions should take longer to answer than demographic questions. Attitude questions have been found to take the longest to answer, as they often require respondents to determine what information is relevant, retrieve it, and then integrate it into an attitude that may not have previously existed (Bassili and Fletcher 1991; Tourangeau et al. 2000; Yan and Tourangeau 2008). In addition, complex attitude questions may take longer to answer than less complex attitude questions (e.g., Yan and Tourangeau 2008). The format of the response task can also strongly affect response time. We expect that open-ended questions will take the longest to answer. Descriptive open-ended questions where respondents answer in their own words are known to be highly burdensome to respondents and to take considerable time to answer. For example, Smyth, Dillman, Christian, and McBride (2009) found that descriptive open-ended questions asked in a web survey took around a minute to answer. By comparison, open-ended questions requiring a numeric response can be answered more quickly because they require less information to be communicated. Couper, Kennedy, Conrad, and Tourangeau (2011) found that numeric open-ended items took from 10 to 20 seconds to answer in a web survey. Questions with closed-ended response formats can be answered more quickly than open-ended questions (Couper and Kreuter 2013), often in only a few seconds (Bassili and Fletcher 1991; Yan and Tourangeau 2008). Within the closed-ended formats, we expect respondents to be able to answer ordinal scale questions more quickly than nominal questions. In nominal questions, respondents have to process each response option individually, whereas processing can be accelerated with ordinal questions with a logical order to the response options. An exception to this is yes/no questions, which should be answered more quickly than other types of questions.

2.2.2 Respondent task complexity. A second set of design features are those that may contribute to the complexity of the respondents' task and therefore may increase response time. One such feature is the reading level of the question. Questions with higher reading levels will be more difficult for respondents to comprehend and process and may create confusion for them, increasing response latency (Yan and Tourangeau 2008). Increased reading level may also make questions more difficult for interviewers to administer. Another feature that may increase response time by increasing respondent task complexity is a mismatch between what is asked in the question stem and what is asked in the response options. Such a mismatch occurs when the question stem implies one type of answer is required while the response options require a different type of response (Fowler 1995; Smyth 2008; Dillman et al. 2014). A simple example is a question that asks, "Are you satisfied with X? Very satisfied, satisfied, dissatisfied, or very dissatisfied." Here, the question stem asks for a "yes" or "no" answer, but the response options require considerably more

information. Questions with mismatches between the stem and the response options are expected to require more processing and possibly add interviewer/respondent interactions (Dillman et al. 2014), both of which will increase the time to answer the question.

Additional question features that affect respondent task complexity and therefore may affect response time are whether the question contains unknown terms or is about a sensitive topic. Unknown terms are terms with which respondents are unlikely to be familiar. For example, this survey asked about the obscure sports of kaninhop and octopush. Examples from other surveys include non-existent legislative acts (e.g., the "Agricultural Trade Act of 1978," Schuman and Presser 1981) and specialized medical terminology (e.g., "myocardial infarction," "angina," and "incontinence" in the 2014 BRFSS questionnaire, Centers for Disease Control and Prevention 2013). Questions that contain unknown terms such as these require additional processing by the respondent at the comprehension stage and as a result should take longer to answer.

Sensitive questions are those asked about topics that respondents may be uneasy or embarrassed to report, out of concern for appearances, possible repercussions, or because the questions are intrusive (Tourangeau et al. 2000). Examples in the literature are questions about illicit drug use (Tourangeau and Smith 1996), abortion (Fu, Darroch, Henshaw, and Kolb 1998), drinking and driving (Dillman and Tarnai 1991), poor college performance (Kreuter, Presser, and Tourangeau 2008), and traffic violations (Bradburn, Sudman, and Associates 1979). These questions may speed or slow response time. Those respondents who have not engaged in the sensitive behavior and those who have but make a quick heuristic decision not to provide an answer that reflects their true behavior can answer quickly. The remaining respondents may engage in multiple cognitive processing steps to retrieve and process relevant information and then decide whether or how much to edit their response, thus increasing response times (Tourangeau and Yan 2007). Given the awkward nature of sensitive questions for conversation, we also expect less interviewer/respondent interaction on sensitive questions than on non-sensitive questions. Thus, we expect the net effect on response time for sensitive questions to be negative; that is, that sensitive questions will be answered more quickly than non-sensitive questions.

2.2.3 Interviewer task complexity. Other features are expected to make the interviewer's task more complex and increase response time. These include question features, visual design features of the interviewer's screen, and indicators of problems during the interview. For example, questions with extra instructions such as "[do not] read response options" or "record answer verbatim" require interviewers to process information on their screen and incorporate it into their interviewing task. These added steps are expected to increase response time, although Couper and Kreuter (2013) surprisingly found that questions with such instructions took less time in an in-person survey. Whether or not a

question has an accompanying probe on the screen also increases the complexity of the interviewer's task and thus may increase response time.

In addition to instructions, we expect that the visual design features of parentheses, questions split over two screens, and questions with emphasis (e.g., bold, underlining, all caps) will increase interviewer task difficulty and thus increase response time. These questions introduce uncertainty about what to read and how to read it and require extra navigational steps. To our knowledge, no previous research has examined these visual design features or their effects on response time.

Finally, survey paradata provide information about other behaviors during an interview that may indicate interviewer task difficulty. In particular, we expect that backing up in a questionnaire makes the interviewer's job more difficult. Backing up may occur because the interviewer inadvertently entered the wrong answer, the respondent changed his/her mind about an answer, or after providing clarification or a definition, the interviewer and respondent came to a different answer. Thus, we include an indicator for whether or not the interviewer backed up at a particular question for a given respondent.

2.2.4 Processing efficiency. There are also features of questionnaires that might cause respondents to stop and think about their responses or help them process more efficiently. For example, items are said to be arranged in a battery when they are presented after a common introduction and with shared response options (Saris and Gallhofer 2014). An example is "On a scale of one to five, where five means you enjoy the activity completely and one means you do not enjoy the activity at all, please tell me how much you enjoy the following leisure activities. First, how about reading? What about cooking? Arts and crafts?" We expect the first question in a battery to take longer because it typically has more words and sets the context for all of the items in the remainder of the battery (Saris and Gallhofer 2014). By the time they get to the later items, we expect respondents to know what they are being asked and the response scale so they can answer subsequent items more quickly.

We also expect extra words in a question—such as definitions and transition statements (e.g., "the next question is going to ask you about how often you've engaged in exercise")—to increase response time because they add additional words for interviewers to read and respondents to process.

Finally, we expect response time to be affected by whether a question is part of a skip pattern where responses to a filter question(s) are used to determine who should answer follow-up questions and who should skip past them. We expect no effect on the filter question in a CATI survey. However, we expect follow-up items (i.e., the subquestions that only some respondents are asked) to have shorter response times because they are generally topically related to the filter question. Relevant information already retrieved for the filter question will be much more available in answering follow-up questions. This is consistent with Tourangeau, Rasinski, and D'Andrade's (1991)

finding that respondents answer attitude questions more quickly if they have previously been asked questions on the same topic, because relevant information is more accessible.

2.3 Respondent Characteristics

In as much as older individuals and those with lower education levels are expected to have lower cognitive and working memory abilities (Krosnick 1991; Narayan and Krosnick 1996; Knauper 1999), they are also expected to have longer response times. Previous research has confirmed this hypothesis in web (Yan and Tourangeau 2008) and CAPI surveys (Couper and Kreuter 2013). In addition, two key skip patterns in the questionnaire examined here are triggered by employment status and computer usage questions. Thus, we include whether or not the respondent is currently employed and whether they have a desktop, laptop, or tablet computer as covariates. We expect that individuals who are employed have greater time constraints than those who are not employed and thus will answer questions more quickly. We also expect that persons who have a computer will answer more quickly, as computer programs and websites frequently request information, perhaps making the question answering process more familiar. We also include a control variable for respondent sex to account for potential overrepresentation of women in phone surveys.

2.4 Interviewer Characteristics

There are few examinations of interviewer characteristics and response time. A consistent finding is that interviewers get faster at conducting interviews over the course of the field period (Olson and Peytchev 2007; Olson and Bilgen 2011). One hypothesized reason for this acceleration is that fluency of administration increases and that interviewer difficulties are greater at the start of the field period. Thus, we expect response time to decrease over the course of the field period. Other interviewer characteristics (e.g., age, sex) do not have sufficient previous empirical research to form expectations, and those that have been examined (e.g., Couper and Kreuter 2013) have not shown consistent findings over subpopulations in the same survey. Although we know that different interviewers recruit different types of respondents (West and Olson 2010), we do not know how fixed characteristics—such as sex or race of interviewers—may be related to response time.

Previous research has shown that experienced and inexperienced interviewers change their pace in different ways over the course of the field period (Olson and Peytchev 2007), obtain different responses to the same questions (e.g., Cleary, Mechanic, and Weiss 1981; Chromy, Eyerman, Odom, Madeline, and Hughes 2005), and are differentially linked to certain error-producing response behaviors, such as acquiescence (Olson and Bilgen 2011). This research suggests that experienced and inexperienced interviewers act differently in an interview.

Relatively unexamined is whether these differences between experienced and inexperienced interviewers occur for all questions in a questionnaire or are concentrated in certain types of questions. In the only analysis of which we are aware, Couper and Kreuter (2013) report inconsistent or nonsensical interaction effects between interviewer experience and question characteristics. We expect experienced interviewers to differ in their question asking and probing behaviors from inexperienced interviewers on questions for which the interviewer has the most latitude, and that this will manifest in shorter time spent on particular types of questions. For example, experienced interviewers may shortcut longer or more difficult questions, or may probe differentially on open-ended, sensitive, or burdensome questions (Fowler and Mangione 1990). Thus, we will also examine whether there are interaction effects between interviewer experience and question features.

3. Data and Methods

3.1 Data

The data for this paper come from the Work and Leisure Today (WLT) survey. The WLT is a landline RDD CATI survey fielded by AbtSRBI between July 31, 2013, and August 28, 2013 (n = 450, AAPOR RR3 = 6.3 percent). The target population for this study was US adults in landline households. Adults were randomly selected within households using the Rizzo method (Rizzo, Brick, and Park 2004). At the time of the survey, an estimated 38 percent of US adults lived in cell-phone-only households (Blumberg and Luke 2013); this group is not represented in this survey. The WLT questionnaire covered topics such as employment status, views on the respondent's employer, leisure activities, computer activities, and demographics. The survey took 15 minutes on average to complete.

Our dependent variable for this analysis is the log-transformed number of seconds spent on each question. The CATI instrument was programmed in the Voxco CATI software system. The Voxco system recorded the time each question started and the duration of time in seconds it took each question to be administered. We calculated the time for administration of each question by taking the difference of start times for subsequent questions for each respondent. As with most response time paradata, the response times for each question are highly skewed. We use two transformations recommended for analyzing response times (Yan and Olson 2013). First, we truncate the distribution of times by replacing all values below and above the first and 99th percentiles with those percentiles, respectively. Second, we take a natural log transformation of all of the times. Before any transformations, the average number of seconds per question is 15.05 seconds (SD = 13.64); after truncating the distribution at the first and 99th percentiles, the average number of seconds per question is

14.77 seconds (SD = 11.78). With a log transformation, the average number of logseconds per question is 2.45 (SD = 0.68).

The primary independent variables for this analysis are characteristics of the 54 questions and CATI screens (see table 2). To obtain the question and screen characteristics, each question and screen was independently coded by two trained graduate student coders (kappas range from 0.85 to 1.00 for the codes examined here), with codes verified by the two authors as master coders. Discrepancies were resolved by the master coders.

The screen and question codes fall into four main categories—Necessary Question Features, Respondent Task Complexity, Interviewer Task Complexity, and Processing Efficiency. Table 2 shows the distribution of each characteristic over each of the 54 survey questions. The necessary question characteristics include the number of words in the question ($\bar{x}$ = 14.56, SD = 12.71), the number of response options ($\bar{x}$ = 3.39, SD = 3.49), the type of question (43 percent behavior, 31 percent attitude/opinion, 26 percent demographic), and the format of the response options (31 percent open-ended numeric, 33 percent closed-ended numeric, 9–15 percent each open-ended text, yes/no, and closed-ended nominal). Each respondent has a respondent-specific counter for the sequential number of the questions that they have been asked, reflecting skip patterns.

The indicators of questions that may increase the task difficulty for respondents are a question's reading level ($\bar{x}$ = 6.6, indicating a grade level between sixth and seventh grade), a mismatch between the question task and the response options (13 percent), questions that are sensitive (13 percent; e.g., counternormative or private topics—being fired from a job, receiving parking or speeding tickets, having sex, looking at "adult" websites, drinking alcohol, income), and questions with unknown terms (4 percent).

The last two groups of question and screen characteristics include those that increase the complexity of the interviewer's task and those that increase or decrease a respondent's processing efficiency. Characteristics of the interviewer's CATI screen that may make the interviewer's task more difficult include instructions for the interviewer (37 percent), the presence of parentheses (9 percent), whether the question is asked over two screens (31 percent), whether there are probes (5.6 percent), and whether there is emphasis in the question (15 percent).We also include a variable created from the paradata that indicates whether the interviewer backed up to a particular question for a respondent (1.1 percent). Those characteristics that affect processing efficiency include whether the question is the first (7.4 percent) or a later question (33 percent) in a battery, whether the question contains definitions (19 percent), whether the question contains a transition statement (13 percent), and whether the question is the first question (the "feeder" question) in a skip pattern (5.6 percent) or a later question in a skip pattern (30 percent).

We examine five respondent characteristics. These respondent characteristics are not intended to be comprehensive of all possible respondent characteristics, but are characteristics that have been previously shown to be associated

Table 2. Descriptive Statistics for Question, Respondent, and Interviewer Characteristics

	n	%/mean	SD
Trimmed number of seconds per question	21,025	14.775	11.797
Log(trimmed number of seconds per question)	21,025	2.448	0.680
Necessary question features			
#Words in question	54	14.556	12.713
# Response options	54	3.389	3.488
Type of question			
Attitude/opinion	17	31.48%	
Behavior	23	42.59%	
Demographics	14	25.93%	
Format of response options from R's point of view			
Open-ended text	5	9.26%	
Open-ended numeric	17	31.48%	
Closed-nominal	6	11.11%	
Closed-ordinal	18	33.33%	
Yes/no	8	14.81%	
Sequence number	54	26.314	
Respondent task complexity			
Question reading level	54	6.635	4.761
Mismatch between question and response options	54	12.96%	
Sensitive question	54	12.96%	
Unknown terms	54	3.70%	
Interviewer task complexity			
Interviewer instructions	54	37.04%	
Parentheses	54	9.26%	
Question asked on two screens	54	31.48%	
Probes on screen	54	5.56%	
Emphasis	54	14.81%	
Backup	21,025	1.07%	
Processing efficiency			
Question in battery			
First question in battery	4	7.41%	
Later questions in battery	18	33.33%	
Not in battery	32	59.26%	
Definitions	54	18.52%	
Transition statement	54	12.96%	
Feeder questions in skip pattern	54	5.56%	
Later questions in skip pattern	54	29.63%	
Respondent characteristics			
Age	450	61.361	16.707
Respondent education = High school degree or less	450	28.67%	
Respondent sex = Female	450	63.78%	
Respondent currently employed	184	40.89%	
Have a laptop, desktop, or tablet computer	350	77.78%	

Table 2. Descriptive Statistics for Question, Respondent, and Interviewer Characteristics *(continued)*

	n	%/mean	SD
Interviewer characteristics			
Interviewer sex = Female	12	54.55%	
Interviewer race = White	9	40.91%	
Interviewer experience = 1 year or more	15	68.18%	
Workload	22	20.45	7.652

with response time or are associated with the total number of questions in this questionnaire due to skip patterns. We include the respondent's age ($\bar{x}$ = 61.4 years), educational level (29 percent have a high school degree or less), sex (64 percent female), employment status (41 percent currently employed), and whether they have a computer (78 percent). Missing data for respondent age were filled in with the mean of the observed respondent ages for the four sex × education cells.

Finally, we include four interviewer characteristics. There are 22 interviewers, 12 of whom (54.5 percent) are female, 9 (41 percent) are white, and 15 (68 percent) have at least one year of experience. The average interviewer workload is 20.45 respondents (min = 5, max = 27).

3.2 Methods

The data have a complex nested structure. Each response time value is nested within respondents, within questions, and within interviewers. With 450 respondents, up to 54 questions per respondent, and 22 interviewers, we have n = 21,025 interviewer-respondent questions in the data set.

We expect that each respondent, interviewer, and question will have a unique effect on response time. To account for this nesting, we estimate a cross-classified random effects model with response times cross-classified by respondents and by questions and with questions and respondents nested within interviewers. Figure 1 displays the data structure visually.

Following notation by Beretvas for a three-level cross-classified model (2010, pp. 330–331), the base model predicts the natural logarithm of response time ($Y_{i(j1,\, j2)k}$) as a function of an overall mean (γ_{0000}) plus a random effect due to the respondent ($u_{0\, j10k}$), a random effect due to the question ($u_{00\, j2k}$), a random effect due to the interviewer (v_{000k}), and a residual term ($e_{i(j1,j2)k}$), where $u_{0\, j10k}$, $u_{00\, j2k}$, and v_{000k} are normally distributed with mean zero and variance $\tau_{u\, j10}$,

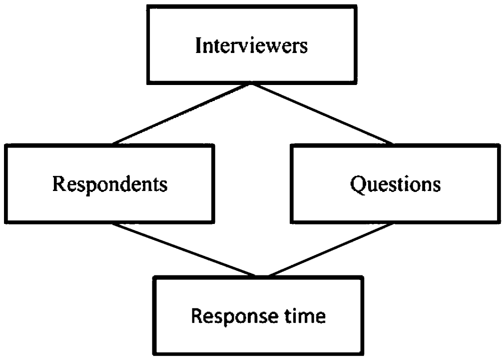

Figure 1. Data Structure of Response Time Cross-Classified by Respondents and Questions and Nested in Interviewers.

$\tau_{u\,j2}$, and τ_{uk}, respectively, and $e_{i(j1,j2)k}$ is normally distributed with mean zero and variance $\sigma_e^{\,2}$ (Beretvas 2010, p. 330):

$$Y_{i(j1,\,j2)k} = \gamma_{0000} + v_{000k} + u_{0\,j10k} + u_{00\,j2k} + e_{i(j1,j2)k}$$

We use the base model to obtain intraclass correlation coefficients to evaluate how much of the variance in log(response time) is due to respondents versus questions versus interviewers. These intraclass correlation coefficients are calculated as

$$\rho_{\text{resp}} = \frac{\hat{\tau}_{uj10}}{\hat{\tau}_{uj10} + \hat{\tau}_{uj2} + \hat{\tau}_{uk} + \hat{\sigma}_e^{\,2}}$$

for the variance due to respondents, and the equation modified with the appropriate variance due to questions in the numerator for the intraclass correlation coefficient due to questions and interviewers (Raudenbush and Bryk 2002).

The base model is then expanded to include question, respondent, and interviewer characteristics:

$$Y_{i(j1,\,j2)k} = \gamma_{0000} + \sum_{m=1}^{p} \beta_m \text{Respondent\_char}_{\,j10} + \sum_{s=1}^{q} \beta_s \text{Question\_char}_{\,j2} + \sum_{t=1}^{r} \beta_t \text{Iwer\_char}_{\,k} + v_{000k} + u_{0j10k} + u_{00j2k} + e_{i(j1,j2)k}$$

We are particularly interested in the relationship between the question characteristics and response time (indicated by the β_s coefficients). Given that there were skip patterns in the questionnaire, different respondents received different sets of questions. All continuous question characteristics (e.g., question order, reading level, number of words, number of response options) are centered at the grand mean. All of the models are estimated using restricted maximum

likelihood estimation in Stata 13.1 xtmixed with random intercepts for questions, respondents, and interviewers (Rabe-Hesketh and Skrondal 2012).

With the exception of the base model, all models include respondent and interviewer characteristics so that we can evaluate the effects of question characteristics net of any potentially confounding respondent characteristics due to skip patterns in the questionnaire. We start with a model containing only respondent and interviewer characteristics. We add the necessary question characteristics as the first set of question characteristics, and control for these characteristics in each subsequent model (preliminary analyses indicated that there were strong suppression effects when these question characteristics were not included).We estimate a model that combines all of the question characteristics, and then estimate a parsimonious model that includes only those question characteristics that were statistically significant at the $p < 0.05$ level in at least one of the previous models. Using findings from the parsimonious model, we then evaluate whether there are statistically significant interactions between interviewer experience and each of the statistically significant question characteristics. Results from the intermediate and interaction models are presented in the appendix; only the final combined model and parsimonious model results are displayed here.

4. Findings

4.1 Base Model

Table 3 shows the results from the base model; that is, the model that contains no covariates. In this model, there are significant variance terms for the question, respondent, and interviewer. This indicates significant variability in response time due to all three sources. The intraclass correlation coefficient for

Table 3. Model Variance Components, Predicting Log(# seconds)

	Square root (Variance)	ICC
Null model		
Interviewer $\sqrt{\tau_{uk}}$	0.123****	0.032
Question $\sqrt{\tau_{uj2}}$	0.505****	0.534
Respondent $\sqrt{\tau_{uj10}}$	0.188****	0.074
Residual $\sqrt{\sigma_e^2}$	0.414****	
Model fit statistics		
Log-likelihood	−12027.609	
AIC	24065.22	

n = 450 respondents, 54 questions, and 22 interviewers. Total n = 21025.
**** $p < 0.0001$; variance components tested using mixture of chi-square distributions.

interviewers is only 0.032, indicating that 3.2 percent of the variance in response time is due to interviewers. The intraclass correlation coefficient due to questions is 0.534 and due to respondents is 0.074, indicating that 53.4 percent of the variance in response time is due to the question and 7.4 percent of the variance in response time is due to respondents. Said another way, roughly seven times the variance in response time in the WLT survey is due to questions versus respondents, and the variance due to questions is more than 15 times that due to respondents. These findings replicate those of Couper and Kreuter (2013), who also found that most of the variability in response time was due to questions rather than respondents or interviewers.

Following Rabe-Hesketh and Skrondal (2012, p. 452), we test whether the random effects model is an improvement over a linear regression using a mixture of three chi-square distributions on 1, 2, and 3 degrees of freedom. The inclusion of the random effects significantly improves the fit of the model ($p <$ 0.0001). Thus, we will estimate random effects for respondent, question, and interviewer in each of our models.

4.2 Respondent and Interviewer Characteristics

Now, we evaluate respondent and interviewer characteristics (Table 4). Not surprisingly, older respondents take longer to answer questions (coef = 0.003, $p < 0.0001$). Respondent sex and education are not statistically associated with response time, but employed persons (coef = − 0.054, $p = 0.008$) and persons with a computer (coef = − 0.056, $p = 0.015$) answer more quickly than their unemployed and computer-less counterparts.
None of the interviewer characteristics are associated with response time except for interview order. Consistent with previous research (Olson and Peytchev 2007), interviewers get faster as they conduct more interviews during the field period (coef = − 0.0025, $p = 0.039$). Inclusion of the respondent and interviewer characteristics reduces the variance due to respondents by 22.3 percent. The variance due to interviewers actually increases by 8.1 percent over the base model, reflecting lack of interpenetration.

4.3 Question Characteristics

Table 5 shows the results of the combined and parsimonious models, including the question, respondent, and interviewer characteristics. Using the AIC as the criterion, the parsimonious model has the best fit out of all the models (lowest AIC). Compared to the base model, we have explained $(0.505^2 - 0.215^2)/0.505^2 = 0.8187$ or about 82 percent of the initial variability in response time due to questions with these question and screen characteristics alone. Also compared to the base model, we have explained $(0.188^2 - 0.165^2)/0.188^2 = 0.230$, or about 23 percent of the variation in response time due to respondents with the five respondent characteristics included in the model. We have

Table 4. Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

Coef. (SE)	
Constant	2.52****
	(0.100)
Respondent characteristics	
Female = 1	−0.32
	(0.018)
Age (centered)	0.003****
	(0.001)
HS degree or less = 1	0.024
	(0.020)
Employed = 1	−0.054**
	(0.201)
Have a computer = 1	−0.056*
	(0.023)
Interviewer characteristics	
1 year+ experience = 1	−0.004
	(0.068)
Female = 1	0.029
	(0.059)
White = 1	0.058
	(0.063)
Interview order	−0.003*
	(0.001)
Random effects	
SD – Interviewer	0.127****
SD – Question	0.505****
SD – Respondent	0.165****
SD – Residual	0.414****
Log-likelihood	−12005.162
AIC	24038.32
Wald chi-square	116.66****

n = 450 respondents, 54 questions, and 22 interviewers. Total n = 21,025.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$
Variance components tested with mixtures of chi-square distributions.

not explained any of the interviewer-related variance; in fact, the interviewer variance component increased with the inclusion of respondent characteristics. Normal quantile plots of the estimated random effects indicate that the assumption of normality for the random question effects holds, with slight deviations

Table 5. Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

	Model 1 Combined	Model 2 Parsimonious
Necessary question features		
Sequential question number	0.001	–
	(0.002)	
Question length	0.021****	0.020****
	(0.006)	(0.003)
# Response options	0.050	0.041****
	(0.041)	(0.011)
Type of question		
Behavior (ref)	–	–
Attitude/opinion	−1.352*	−1.080***
	(0.478)	(0.317)
Demographic	−0.331	−0.292**
	(0.137)	(0.108)
Format of response options		
Open-ended text (ref)	–	–
Open-ended numeric	−0.276	−0.310*
	(0.178)	(0.171)
Closed-nominal	−0.617	−0.626****
	(0.384)	(0.143)
Closed-ordinal	0.447	0.298
	0.377)	(0.306)
Yes/No	−0.811*	−0.844****
	(0.399)	(0.131)
Respondent task complexity		
Question reading level	0.027	0.027****
	(0.014)	(0.007)
Mismatch between q'n and response options	−0.182	–
	(0.145)	
Sensitive question	−0.416**	−0.337**
	(0.149)	(0.122)
Unknown terms	0.166	–
	(0.301)	
Interviewer task complexity		–
Interviewer instructions	−0.079	
	(0.230)	
Parentheses	0.057	
	(0.192)	
	–	
Question asked on two screens	0.209	
	(0.367)	–
Probes on screen	−0.061	
	(0.371)	–
Emphasis	−0.210	
	(0.150)	–
Backup at question	0.120****	0.120****
	(0.028)	(0.028)

Table 5. Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds) (*continued*)

	Model 1 Combined	Model 2 Parsimonious
Processing efficiency		
First question in battery	0.079	
	(0.201)	
Later questions in battery	0.072	
	(0.157)	–
Definitions	0.209	0.212*
	(0.145)	(0.101)
Transition statement	0.074	
	(0.141)	–
Feeder questions in skip pattern	0.019	
	(0.200)	–
Later questions in skip pattern	0.084	
	(0.103)	–
Random effects		
SD – Interviewer	0.127****	0.127****
SD – Question	0.232****	0.215****
SD – Respondent	0.165****	0.165****
SD – Residual	0.414****	0.414****
Log-likelihood	−12007.353	−11972.271
Wald chi-square	226.33****	381.57****
AIC	24048.16	23996.54

n = 450 respondents and 54 questions. Total respondents × questions n = 21,025. All models include respondent and interviewer characteristics. Variance components tested with mixtures of chi-square distributions.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

from normality at the tails for the interviewer and respondent random effects in the parsimonious model.[1] Given that the parsimonious model is the best-fitting model and the potential multicollinearity of question characteristics, we will discuss only the parsimonious model results.

1. Normal quantile (Q-Q) plots visually display quantiles of two distributions against each other, allowing one to assess whether the distributions match. In this case, the estimated random effects were graphed against a normal distribution to provide a visual check of the normality assumption.

Among the necessary question feature covariates, as expected, questions with more words (coef. = 0.020, $p < 0.0001$) and more response options take longer to administer (coef. = 0.041, $p < 0.0001$). Open-ended numeric (coef = −0.310, $p < 0.05$), closed ended nominal (coef = −0.626, $p < 0.01$), and yes/no questions (coef. = −0.844, $p < 0.0001$) take less time to administer than open-ended text questions.[2] There is not a significant association between the sequential number of the question in the questionnaire and response time (this coefficient is statistically significant and negative with respondent characteristics excluded). Additionally, both attitude (coef = −1.348, $p < 0.0001$) and demographic questions (coef = −0.424, $p = 0.001$) are answered more quickly than behavioral questions.

The next set of predictors includes question characteristics that may affect task complexity for the respondent. These include the question's reading level, whether there is a mismatch between the question stem and the response options, whether the question is sensitive, and whether it contains unknown terms. Questions that are more difficult to read take more time to answer (coef. = 0.027, $p < 0.0001$), and sensitive questions take less time to answer (coef. = −0.337, $p < 0.01$).

The third set of variables in table 5 includes indicators of question or screen characteristics that make the interviewer's task more complex. To our surprise, only one of these characteristics is statistically different from zero. As expected, response time increases when there is a backup (coef = 0.120, $p < 0.0001$).

The fourth set of variables includes question and screen characteristics that may affect the efficiency of question processing. Of these characteristics, only the presence of definitions on the screen is statistically different from zero. Questions with definitions take longer to administer than questions without definitions (coef = 0.212, $p < 0.05$). Whether the question appears in a battery or in a skip pattern is not associated with response time, nor is whether a transition statement appears on the screen. We also included a counter of the number of an item in a battery and did not find an effect.

4.4 Interactions with Interviewer Experience

The overall effect of interviewer experience is not statistically different from zero. That is, experienced and inexperienced interviewers do not differ in their administration time for each question overall. However, we would expect that experienced and inexperienced interviewers would administer certain types

2. In post-hoc tests, closed-ended nominal are significantly faster than closed-ended ordinal ($\chi_1^2 = 10.60$; $p = 0.0011$) but not significantly different from yes/no questions ($\chi_1^2 = 2.94$; $p = 0.09$). Closed-ended ordinal items take significantly longer than yes/no questions ($\chi_1^2 = 15{:}48$; $p = 0.0001$).

of questions differently. In particular, we expect that experienced interviewers would take shortcuts on longer, sensitive, or more burdensome questions.

There are no noticeable differences for questions with different numbers of response options; for attitudinal, behavior, and demographic questions; for questions with different reading levels; or for sensitive questions. We found statistically significant interaction terms between interviewer experience and question length, response option format, and the presence of definitions. In each instance, the difference in response time for these question characteristics between experienced and inexperienced interviewers was modest (around 1 second/question difference). The findings are summarized in table 6 and presented in the online appendix. The question length interaction term worsened model fit, as did the definition interaction term once the response option format interaction term was included in the model (all determined using AIC criterion). In general, experienced interviewers take longer (about 1.4 seconds per question) to administer open-ended text questions (perhaps indicating greater probing abilities), but slightly less time to administer yes/no questions (about 0.40 seconds) than inexperienced interviewers.

5. Conclusion and Discussion

In this paper, we evaluated the relationship between question features and response time in a CATI survey. We examined both traditional question features and those that reflect the visual design of the question within the CATI instrument. We have three main findings. First, we were able to account for about 80 percent of the variance in response time due to questions with the question characteristics included here. The characteristics of questions that had the biggest influence were the necessary question features and those that affect respondent task complexity. To our surprise, none of the visual design features related to the interviewer's task had a statistically significant effect on response time. Second, respondent age, employment status, and computer use were significantly associated with response time, but education unexpectedly was not. These respondent characteristics accounted for roughly one-quarter of the variance in response time associated with respondents. Finally, although there were no differences overall, we found that experienced interviewers administered yes/no questions more quickly than inexperienced interviewers and took more time on open-ended questions.

It is reassuring to replicate findings from previous research (and common sense!) that questions with more words take longer to administer. In contrast to previous research, we found that attitudinal and demographic questions are answered more quickly than behavioral questions. This could be because the attitudinal questions in this particular survey are questions for which the respondent readily has an answer, whereas the behavioral questions require more comprehension and retrieval effort. Furthermore, the difference between

Table 6. Summary of Predictions and Findings

Question characteristics	Prediction	Finding	Vary with interviewer experience?
Necessary question features			
Question sequence #	−	None	
Length of the question	+	+	Yes
# of response options	+	+	No
Type of question	attitude + behavior (ref) demos −	attitude − behavior (ref) demos −	No
Response options format	open text (ref) open numeric − closed nominal − closed ordinal − yes/no −	open text (ref) open numeric − closed nominal − closed ordinal none yes/no −	Yes
Respondent task complexity			
Question reading levels	+	+	No
Mismatched between question and response options	+	None	
Unknown terms	+	None	
Sensitive questions	−	−	No
Interviewer task complexity			
Interviewer instructions	+	None	
Probe on screen	+	None	
Use of parentheses in question	+	None	
Question asked over two screens	+	None	
Question has visual emphasis	+	None	
Backing up	+	+	
Processing efficiency			
First question in battery	+	None	
Later question in battery	−	None	
Questions with definitions	+	+	Yes
Feeder question	None	None	
Follow-up questions in a skip	−	None	

(ref) indicates the reference group, + indicates increased response time, − indicates decreased response time.

attitude and behavior questions became apparent only after accounting for the reading level of the question, indicating differences in the difficulty of the two question types.

This paper makes several new contributions to existing literature, replicating and expanding beyond the work of Yan and Tourangeau (2008) and Couper

and Kreuter (2013). First, we were able to examine characteristics not previously considered in a systematic cross-questionnaire evaluation of response time. We found that sensitive questions are answered more quickly than nonsensitive questions, that the reading level of questions is positively associated with response time, and that questions with definitions take longer to answer than questions without definitions. We also were surprised that we did not see gains in response times for questions that appear in a battery.

Second, we show that various visual features of the CATI instrument do not have an effect on response time. This finding tentatively suggests that visual design may not have as large an effect in CATI surveys as it does in selfadministered surveys. It is possible that through training and practice with the instrument, interviewers will learn to overcome poor visual design in CATI surveys. Additionally, it may be that visual design affects interviewer administration time of survey questions, but the latent paradata timers used in this research obscure those differences because they contain respondent answering time and additional interviewer-respondent interaction time.

Third, this analysis provides a more direct examination of differences between experienced and inexperienced interviewers on different types of questions. Interestingly, experienced interviewers take longer on open-ended questions, consistent with having better probing behavior (e.g., Fowler and Mangione 1990). But they also go slightly more quickly on yes/no questions, a finding that may explain the increased levels of acquiescence (from lack of deeply processing these yes/no questions; Krosnick and Presser 2010) for experienced interviewers found by Olson and Bilgen (2011).

This paper has limitations. First, although we were able to identify a number of question characteristics that are associated with response time, we did not explicitly link response time to a measure of data quality. For example, it is unclear whether responding to sensitive questions more quickly than nonsensitive questions indicates better or worse data quality on those questions. Similarly, although we found that questions with definitions took longer to answer, we do not know whether the respondent successfully integrates the information in the definition into their response. Additional research is needed to examine other data quality outcomes as well as the interviewer and respondent interactions that occur during the question-and-answer process to ascertain whether these changes in response times might be linked with better or worse data quality. Second, the survey used a landline sample of US adults. Thus, the respondent pool was much older and more educated than all adults in the United States, and persons in mobile-only households are not represented. Additionally, we do not know from this research how these characteristics would operate in a cell phone interview. Third, although there are variations in question characteristics, they were not experimentally assigned. Thus, we cannot evaluate the effects of some question characteristics. For example, although the questionnaire contained a number of items asked using a scale, we cannot evaluate different types of scales (e.g., unipolar versus bipolar, endpoint

labeled versus fully labeled) in this survey. Fourth, although the survey was administered by an organization with a great deal of experience conducting CATI surveys, the Voxco CATI system was new for the organization, and the interviewers were still learning it. Nevertheless, interviewers accounted for only 3.2 percent of the variance in response times, a magnitude in line with Couper and Kreuter's findings in a face-to-face survey. Finally, the response time measures collected by paradata are from "latent timers" (Mulligan et al. 2003), measuring the time from which a question appeared on an interviewer's screen to the time that they advanced to the next question. Thus, we cannot disentangle whether a longer time spent on a question is due to the interviewer taking more time to administer a question, the respondent requiring more time to respond, or deviations from a question-answer-neutral feedback sequence (Maynard and Schaeffer 2002).

Survey researchers often lament the shortcomings of respondents. It is easy to believe that the problems in a questionnaire are due to respondents hurrying and not paying close enough attention. In fact, we have even devised ways to "test" respondents to see if they are paying attention, such as by inserting reverse-worded questions in batteries (e.g., Miller and Baker-Prewitt 2009). However, our findings suggest that the shortcomings of the questionnaire instrument itself may be more to blame than respondents. Question characteristics accounted for seven times as much variance in response time as respondent characteristics. Some of the characteristics that have the most impact on response time are necessary because of the survey topic and the type of data needed. Examples include whether one is measuring an attitude, behavior, or demographic characteristic or which response format is used. But with other characteristics, it is possible to make positive changes. Our findings suggest that writing questions that have lower reading levels, fewer words and response options, and minimizing (or where possible simplifying) definitions are examples. Likewise, reducing sensitive questions or asking them in a context that makes them less sensitive are questionnaire design strategies that might improve the question/answer process. While these findings put the onus for achieving high-quality data squarely on questionnaire designers, it is encouraging to learn that the characteristics that make the most difference in response time are question characteristics that are within the surveyor's control.

Appendix

Table A1. Full Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

	Necessary question features	Respondent task complexity	Interviewer task complexity	Processing efficiency	Combined	Parsimonious
Necessary question features						
Sequential question number	0.001	0.001	0.001	0.001	0.001	–
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
Question length	0.023****	0.020****	0.022****	0.028****	0.021****	0.020****
	(0.004)	(0.003)	(0.005)	(0.005)	(0.006)	(0.003)
# Response options	0.039*	0.034***	0.034	0.051****	0.050	0.041****
	(0.014)	(0.011)	(0.041)	(0.013)	(0.041)	(0.011)
Type of question						
Behavior (ref)	–	–	–	–	–	–
Attitude/opinion	−0.626	−1.348****	−0.781	−0.418	−1.352*	−1.080***
	(0.357)	(0.309)	(0.498)	(0.334)	(0.478)	(0.317)
Demographic	−0.318*	−0.424****	−0.301*	−0.099	−0.331	−0.292**
	(0.126)	(0.098)	(0.137)	(0.133)	(0.137)	(0.108)
Format of response options						
Open-ended text(ref)	–	–	–	–	–	–
Open-ended numeric	−0.385*	−0.297*	−0.415*	−0.451**	−0.276	−0.310*
	(0.171)	(0.126)	(0.207)	(0.158)	(0.178)	(0.171)

Continued

Table A1. Full Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

	Necessary question features	Respondent task complexity	Interviewer task complexity	Processing efficiency	Combined	Parsimonious
Closed-nominal	−0.607**	−0.546****	−0.672	−0.750****	−0.617	−0.626****
	(0.195)	(0.142)	(0.389)	(0.188)	(0.384)	(0.143)
Closed-ordinal	−0.079	0.456	0.024	−0.318	0.447	0.298
	(0.356)	(0.300)	(0.414)	(0.327)	(0.377)	(0.306)
Yes/no	−0.850****	−0.827****	−0.902*	−0.956****	−0.811*	−0.844****
	(0.183)	(0.134)	(0.399)	(0.171)	(0.399)	(0.131)
Respondent task complexity						
Question reading level		0.025***			0.027	0.027****
		(0.009)			(0.014)	(0.007)
Mismatch between q'n and response options		−0.123			−0.182	–
		(0.100)			(0.145)	
Sensitive question		−0.456***			−0.416**	−0.337**
		(0.120)			(0.149)	(0.122)
Unknown terms		0.319			0.166	–
		(0.211)			(0.301)	
Interviewer task complexity						
Interviewer instructions			−0.104		−0.079	–
			(0.257)		(0.230)	
Parentheses			0.115		0.057	–
			(0.183)		(0.192)	
Question asked on two screens			0.069		0.209	–
			(0.390)		(0.367)	

Probes on screen			0.170		−0.061	–
			(0.392)		(0.371)	
Emphasis			0.028		−0.210	–
			(0.172)		(0.150)	
Backup at question			0.120****		0.120****	0.120****
			(0.028)		(0.028)	(0.028)
Processing efficiency						
First question in battery				−0.007	0.079	–
				(0.179)	(0.201)	
Later questions in battery				0.083	0.072	–
				(0.158)	(0.157)	
Definitions				0.459****	0.209	0.212*
				(0.117)	(0.145)	(0.101)
Transition statement				−0.069	0.074	–
				(0.146)	(0.141)	
Feeder questions in skip pattern				−0.020	0.019	–
				(0.188)	(0.200)	
Later questions in skip pattern				−0.073	0.084	–
				(0.090)	(0.103)	
Respondent characteristics						
Female = 1	−0.032	−0.032	−0.032	−0.032	−0.032	−0.032
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Age	0.003****	0.003****	0.003****	0.003****	0.003****	0.003****
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.001)
HS degree or less = 1	0.024	0.024	0.024	0.024	0.024	0.024
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)

Continued

Table A1. Full Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

	Necessary question features	Respondent task complexity	Interviewer task complexity	Processing efficiency	Combined	Parsimonious
Employed = 1	−0.062**	−0.058*	−0.062*	−0.057*	−0.057*	−0.054**
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.020)
Has computer = 1	−0.058*	−0.057*	−0.057*	−0.057*	−0.057*	−0.056*
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
Interviewer characteristics						
1 year+ experience = 1	−0.004	−0.004	−0.004	−0.004	−0.004	−0.004
	(0.068)	(0.068)	(0.068)	(0.068)	(0.068)	(0.068)
Female = 1	0.029	0.029	0.029	0.029	0.029	0.028
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
White = 1	0.057	0.057	0.057	0.057	0.058	0.058
	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)	(0.062)
Interview order	−0.003*	−0.003*	−0.003*	−0.003*	−0.003*	−0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	3.160****	3.259****	3.184****	3.084****	3.124****	3.141****
	(0.359)	(0.139)	(0.355)	(0.159)	(0.302)	(0.131)
Random effects						
SD – Interviewer	0.127****	0.127****	0.127****	0.127****	0.127****	0.127****
SD – Question	0.307****	0.221****	0.322****	0.272****	0.232****	0.215****
SD – Respondent	0.166****	0.166****	0.165****	0.165****	0.165****	0.165****
SD – Residual	0.414****	0.414****	0.414****	0.414****	0.414****	0.414****
ICC – Interviewer	0.052	0.061	0.051	0.055	0.060	0.062
ICC – Question	0.305	0.185	0.326	0.257	0.200	0.177

ICC – Respondent	0.088	0.104	0.085	0.094	0.101	0.104
Log-likelihood	−11994.501	−11985.087	−11990.345	−11992.515	−12007.353	−11972.271
Wald chi-square	213.96****	350.50****	224.82****	258.55****	226.33****	381.57****
AIC	24035	24024.17	24038.69	24043.03	24048.16	23996.54

Note.—$n = 450$ respondents and 54 questions. Total respondents × questions $n = 21025$. Variance components tested with mixtures of chi-square distributions.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.
****$p < 0.0001$.

Table A2. Full Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds)

	Interaction with question length	Interaction with question format	Interaction with definitions	Interaction with question format and definition
Necessary question features				
Question length	0.021****	0.020****	0.020****	0.020****
	(0.003)	(0.003)	(0.003)	(0.003)
# Response options	0.041****	0.041****	0.041****	0.041****
	(0.011)	(0.011)	(0.011)	(0.011)
Type of question				
Behavior (ref)	–	–	–	–
Attitude/opinion	−1.080***	−1.080***	−1.080***	−1.080***
	(0.317)	(0.317)	(0.317)	(0.317)
Demographics	−0.292**	−0.292**	−0.292**	−0.292**
	(0.108)	(0.108)	(0.108)	(0.108)
Format of response options				
Open-ended text(ref)	–	–	–	–
Open-ended numeric	−0.310*	−0.237	−0.310*	−0.246*
	(0.123)	(0.125)	(0.123)	(0.125)
Closed-nominal	−0.627****	−0.524****	−0.627****	−0.531****
	(0.143)	(0.145)	(0.143)	(0.145)
Closed-ordinal	0.299	0.378	0.298	0.373
	(0.305)	(0.306)	(0.305)	(0.306)
Yes/no	−0.844****	−0.720****	−0.844****	−0.729****
	(0.131)	(0.133)	(0.131)	(0.133)
Respondent task complexity				
Question reading level	0.027****	0.027****	0.027****	0.027****
	(0.007)	(0.007)	(0.007)	(0.007)

Sensitive question	−0.337**	−0.337**	−0.337**	−0.337**
	(0.122)	(0.122)	(0.122)	(0.122)
Interviewer task complexity				
Backup at question	0.121****	0.121****	0.119****	0.121****
	(0.028)	(0.028)	(0.028)	(0.028)
Processing efficiency				
Definitions	0.212*	0.212*	0.257*	0.246*
	(0.101)	(0.101)	(0.101)	(0.101)
Respondent characteristics				
Female = 1	−0.032	−0.032	−0.032	−0.032
	(0.018)	(0.018)	(0.018)	(0.018)
Age	0.003****	0.003****	0.003****	0.003****
	(0.001)	(0.001)	(0.001)	(0.001)
HS degree or less = 1	0.024	0.024	0.024	0.024
	(0.020)	(0.020)	(0.020)	(0.020)
Employed = 1	−0.054**	−0.054**	−0.054**	−0.054**
	(0.020)	(0.020)	(0.020)	(0.020)
Has computer = 1	−0.056*	−0.056*	−0.056*	−0.056*
	(0.023)	(0.023)	(0.023)	(0.023)
Interviewer characteristics				
1 year+ experience = 1	−0.004	0.103	0.008	0.103
	(0.068)	(0.071)	(0.068)	(0.071)
Female = 1	0.028	0.029	0.029	0.029
	(0.059)	(0.059)	(0.059)	(0.059)

Continued

Table A2. Full Model Coefficients and Standard Error (in parentheses) Predicting Log(# seconds) (*continued*)

	Interaction with question length	Interaction with question format	Interaction with definitions	Interaction with question format and definition
White = 1	0.058	0.058	0.058	0.058
	(0.062)	(0.063)	(0.063)	(0.063)
Interview order	−0.003*	−0.003*	−0.003*	−0.003*
	(0.001)	(0.001)	(0.001)	(0.001)
Interaction effects: Interviewer 1 year + experience *				
Question length	−0.001*			
	(0.0005)			
Open-ended text(ref)		–		–
Open-ended numeric		−0.099****		−0.087****
		(0.023)		(0.024)
Closed-nominal		−0.138****		−0.129****
		(0.029)		(0.029)
Closed-ordinal		−0.107****		−0.100****
		(0.024)		(0.024)
Yes/no		−0.168****		−0.155****
		(0.026)		(0.027)
Definitions			−0.061****	−0.046**
			(0.016)	(0.007)
Constant	3.142****	3.062****	3.141****	3.062****
	(0.131)	(0.132)	(0.131)	(0.132)
Random effects				
SD – Interviewer	0.127****	0.127****	0.127****	0.127****
SD – Question	0.215****	0.215****	0.215****	0.215****

SD – Respondent	0.165****	0.165****	0.165****	0.165****
SD – Residual	0.414****	0.414****	0.414****	0.414****
Log-likelihood	−11975.682	−11962.605	−11968.669	−11962.07
Wald chi-square	388.22****	424.95****	395.33****	432.50****
AIC	24005.36	23985.21	23991.34	23986.14

Note.—$n = 450$ respondents and 54 questions. Total respondents × *questions $n = 21025$. Variance components tested with mixtures of chi-square distributions.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.
****$p < 0.0001$.

References

Bassili, J. N. (1996), "The How and Why of Response Latency Measurement in Telephone Surveys," in Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, eds. N. Schwarz, and S. Sudman, pp. 319–346, San Francisco: Jossey-Bass.

Bassili, J. N., and J. F. Fletcher (1991), "Response Time Measurement in Survey Research: A Method for CATI and a New Look at Nonattitudes," Public Opinion Quarterly, 55, 331–346.

Beretvas, S. N. (2010), "Cross-Classified and Multiple-Membership Models," in Handbook of Advanced Multilevel Analysis, eds. J. J. Hox, and J. K. Roberts, pp. 313–334, New York: Routledge.

Blumberg, S. J., and J. V. Luke (2013), "Wireless Substitution: Early Release Estimates from the National Health Interview Survey, January-June 2013," retrieved from http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201312.pdf.

Bradburn, N., S. Sudman, and Associates (1979), Improving Interview Method and Questionnaire Design, San Francisco: Jossey-Bass.

Bradburn, N., S. Sudman, and B. Wansink (2004), Asking Questions: The Definitive Guide to Questionnaire Design for Market Research, Political Polls, and Social and Health Questionnaires, San Francisco: Jossey Bass. Centers for Disease Control and Prevention (2013), 2014 Behavioral Risk Factor Surveillance System Questionnaire, Retrieved from http://www.cdc.gov/brfss/questionnaires/pdf-ques/2014_BRFSS.pdf.

Christian, L. M., and D. A. Dillman (2004), "The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions," Public Opinion Quarterly, 68 (1), 58–81.

Chromy, J. R., J. Eyerman, D. Odom, M. E. Madeline, and A. Hughes (2005), "Association between Interviewer Experience and Substance-Use Prevalence Rates in NSDUH," in Evaluating and Improving Methods Used in the National Survey on Drug Use and Health, DHHS Publication No. SMA 05-4044, Methodology Series M-5, eds. Joel Kennet, and Joseph Gfroerer, pp. 59–87, Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Cleary, P. D., D. Mechanic, and N. Weiss (1981), "The Effect of Interviewer Characteristics on Responses to a Mental Health Interview," Journal of Health and Social Behavior, 22, 183–93. Couper, M. (1998), "Measuring Survey Quality in a CASIC Environment," Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 41–49.

Couper, M. P., and F. Kreuter (2013), "Using Paradata to Explore Item-Level Response Times in Surveys," Journal of the Royal Statistical Society Series A: Statistics in Society, 176 (Part 1), 271–286.

Couper, M. P., C. Kennedy, F. G. Conrad, and R. Tourangeau (2011), "Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys," Journal of Official Statistics, 27 (1), 65–85.

de Leeuw, E. (1992), Data Quality in Mail, Face to Face, and Telephone Surveys, Amsterdam: TT-Publikates.

de Leeuw, E. (2005), "To Mix or Not to Mix Data Collection Modes in Surveys," Journal of Official Statistics, 21, 233–255.

Dillman, D. A. (1991), "The Design and Administration of Mail Surveys," Annual Review of Sociology, 17, 225–249.

Dillman, D. A., and J. Tarnai (1991), "Mode effects of Cognitively Designed Recall Questions: A Comparison of Answers to Telephone and Mail Surveys," in Measurement Errors in Surveys, eds. Paul Biemer et al., pp. 73–93, New York: JohnWiley.

Dillman, Don A., J. Smyth, and L. M. Christian (2014), Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method, Hoboken, NJ: JohnWiley and Sons, Inc.

Edwards, B., S. Schneider, and P. D. Brick (2008), "Visual Elements of Questionnaire Design: Experiments with a CATI Establishment Survey," in Advances in Telephone Survey Methodology, eds. J. M. Lepkowski, C. Tucker, and J. M. Brick, et al., pp. 276–296, Hoboken, NJ: John Wiley & Sons. Fowler, F. J. (1995), Improving Survey Questions: Design and Evaluation, Thousand Oaks, CA: Sage.

Fowler, F. J., and T. W. Mangione (1990), Standardized Survey Interviewing: Minimizing Interviewer-Related Error, Newbury Park, CA: Sage.

Fu, H., J. E. Darroch, S. K. Henshaw, and E. Kolb (1998), "Measuring the Extent of Abortion Underreporting in the 1995 National Survey of Family Growth," Family Planning Perspectives, 30, 128–133 & 138.

Jenkins, C., and D. A. Dillman (1997), "Towards a Theory of Self-Administered Questionnaire Design," in Survey Measurement and Process Quality, eds. L. E. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, pp. 165–196, New York: Wiley-Interscience.

Johnson, M. (2004), "Timepieces: Components of Survey Question Response Latencies," Political Psychology, 25, 679–702.

Knauper, B. (1999), "Age and Response Order Effects," Public Opinion Quarterly, 63, 347–370.

Kreuter, F., S. Presser, and R. Tourangeau (2008), "Social Desirability Bias in CATI, IVR and Web Surveys," Public Opinion Quarterly, 72, 847–865.

Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," Applied Cognitive Psychology, 5, 213–236.

Krosnick, J. A., and S. Presser (2010), "Question and Questionnaire Design," in Handbook of Survey Research (2nd ed.), pp. 263–314, Bingley, UK: Emerald House Publishing.

Malhotra, N. (2008), "Completion Time and Response Order Effects in Web Surveys," Public Opinion Quarterly, 72, 914–943.

Maynard, D. W., and N. C. Schaeffer (2002), "Standardization and Its Discontents," in Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, eds. D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer,, and J. van der Zouwen, pp. 3–45, New York: JohnWiley & Sons, Inc.

Miller, J., and J. Baker-Prewitt (2009), "Beyond 'Trapping' the Undesirable Panelist: The Use of Red Herrings to Reduce Satisficing," paper presented at the 2009 CASRO Panel Quality Conference, retrieved from http://www.survey4.burke.com/Library/Conference/Beyond%20Trapping%20the%20Undesirable%20Panelist_FINAL.pdf

Mulligan, K., J. T. Grant, S. T.Mockabee, and J. Q.Monson (2003), "Response LatencyMethodology for Survey Research:Measurement andModeling Strategies," Political Analysis, 11, 289–301.

Narayan, S., and J. A. Krosnick (1996), "Education Moderates Some Response Effects in Attitude Measurement," Public Opinion Quarterly, 60, 58–88.

Olson, K., and I. Bilgen (2011), "The Role of Interviewer Experience on Acquiescence," Public Opinion Quarterly, 75, 99–114.

Olson, K., and B. Parkhurst (2013), "Collecting Paradata for Measurement Error Evaluations," in Improving Surveys with Paradata: Analytic Uses of Process Information, ed. F. Kreuter, pp. 43–72, Hoboken, NJ: JohnWiley & Sons.

Olson, K., and A. Peytchev (2007), "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes," Public Opinion Quarterly, 71, 273–286.

Rabe-Hesketh, S., and A. Skrondal (2012), Multilevel and Longitudinal Modeling Using Stata, Third Edition, Volume I: Continuous Responses, College Station, TX: Stata Press.

Raudenbush, S. W., and A. S. Bryk (2002), Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.), Newbury Park, CA: Sage.

Rizzo, L., J. M. Brick, and I. Park (2004), "A Minimally Intrusive Method for Sampling Persons in Random Digit Dial Surveys," Public Opinion Quarterly, 68, 267–274.

Saris, W. E., and I. N. Gallhofer (eds.) (2014), Survey Items in Batteries, in Design, Evaluation, and Analysis of Questionnaires for Survey Research (2nd ed.), Hoboken, NJ: John Wiley & Sons, doi:10.1002/9781118634646.

Schuman, H., and S. Presser (1981), Questions and Answers In Attitude Surveys: Experiments on Question form, Wording, and Context, New York: Academic Press.

Smyth, J. D. (2008, May 15–18). "Unresolved Issues in Multiple-Answer Questions," paper presented at the American Association for Public Opinion Research, New Orleans, LA.

Smyth, J. D., D. A. Dillman, L. M. Christian, and M. McBride (2009), "Open-Ended Questions In Web Surveys: Can Increasing the Size of Answer Spaces and Providing Extra Verbal Instructions Improve Response Quality?" Public Opinion Quarterly, 73, 325–337.

Tourangeau, R., and T.W. Smith (1996), "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context," Public Opinion Quarterly, 60, 275–304.

Tourangeau, R., and T. Yan (2007), "Sensitive Questions in Surveys," Psychological Bulletin, 133, 859–883.

Tourangeau, R., K. Rasinski, and R. D'Andrade (1991), "Attitude Structure and Belief Accessibility," Journal of Experimental Social Psychology, 27, 48–75.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000), The Psychology of Survey Response, New York: Cambridge University Press.

West, B., and K. Olson (2010), "How Much of Interviewer Variance is Really Nonresponse Error Variance?" Public Opinion Quarterly, 74, 1004–1026.

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, pp. 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 Response Times," Applied Cognitive Psychology, 22, 51–68.