

2018

The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews

Kristen Olson

University of Nebraska-Lincoln, kolson5@unl.edu


Jolene Smyth

University of Nebraska-Lincoln, jsmyth2@unl.edu

Amanda Ganshert

University of Nebraska-Lincoln, aganshert@unl.edu

Follow this and additional works at: <http://digitalcommons.unl.edu/sociologyfacpub>

 Part of the [Family, Life Course, and Society Commons](#), [Quantitative, Qualitative, Comparative, and Historical Methodologies Commons](#), [Social Psychology and Interaction Commons](#), and the [Social Statistics Commons](#)

Olson, Kristen; Smyth, Jolene; and Ganshert, Amanda, "The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews" (2018). *Sociology Department, Faculty Publications*. 541.

<http://digitalcommons.unl.edu/sociologyfacpub/541>

This Article is brought to you for free and open access by the Sociology, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Sociology Department, Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Published in *Journal of Survey Statistics and Methodology* (2018), 34p.

doi: 10.1093/jssam/smy006

Copyright © 2018 Kristen Olson, Jolene D. Smyth, and Amanda Ganshert.

Published by Oxford University Press on behalf of the American Association for Public Opinion Research. Used by permission.

An earlier version of this paper was presented at the annual conference of the American Association for Public Opinion Research, in Austin, Texas, in May 2016.

The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews

Kristen Olson, Jolene D. Smyth, and Amanda Ganshert

Kristen Olson is the Leland J. and Dorothy H. Olson Associate Professor and Vice Chair of the Department of Sociology at the University of Nebraska–Lincoln, 703 Oldfather Hall, Lincoln, NE 68588-0324.

Jolene D. Smyth is an associate professor in the Department of Sociology and the Director of the Bureau of Sociological Research at the University of Nebraska–Lincoln.

Amanda Ganshert is a project manager at the Bureau of Sociological Research at the University of Nebraska–Lincoln.

Corresponding author — Kristen Olson, Department of Sociology, University of Nebraska- Lincoln. 703 Oldfather Hall, Lincoln, NE 68588; email kolson5@unl.edu

Abstract

In a standardized telephone interview, respondents ideally are able to provide an answer that easily fits the response task. Deviations from this ideal question answering behavior are behavioral manifestations of breakdowns in the cognitive response process and partially reveal mechanisms underlying measurement error, but little is known about what question characteristics or types of respondents are associated with what types of deviations. Evaluations of question problems tend to look at one question characteristic at a time; yet questions are comprised of multiple characteristics, some of which are easier to experimentally manipulate (e.g., presence of a definition) than others (e.g., attitude versus behavior). All of these characteristics can affect how respondents answer questions. Using a landline telephone interview, we use cross-classified random effects logistic regression models to simultaneously evaluate the effects of multiple question and respondent characteristics on six different respondent behaviors. We find that most of the variability in these respondent answering behaviors is associated with the questions rather than the respondents themselves. Question characteristics that affect the comprehension and mapping stages of the cognitive response process are consistently associated

with answering behaviors, whereas attitude questions do not consistently differ from behavioral questions. We also find that sensitive questions are more likely to yield adequate answers and fewer problems in reporting or clarification requests than nonsensitive questions. Additionally, older respondents are less likely to answer adequately. Our findings suggest that survey designers should focus on questionnaire features related to comprehension and mapping to minimize interactional and data quality problems in surveys and should train interviewers on how to resolve these reporting problems.

Keywords: Interviewer-respondent interaction, Question features, Respondent behaviors, Telephone surveys

1. Introduction

Survey questionnaire designers try to write questions that respondents can ideally answer without follow-up and with responses that easily fit into the response task or response categories (Fowler and Mangione 1990; Blair and Srinath 2008). But respondents often deviate from this ideal. First, respondents may have to *request clarification* about a question before they can answer it. Second, they may provide one of several types of substantive answers that indicate that they cannot easily accomplish the task required by the question, that is, a *problematic substantive response*. For example, respondents may convey uncertainty by qualifying their answers with terms such as “probably” or “I guess,” qualifiers previously shown to be associated with measurement errors (Dykema, Lepkowski, and Blixt 1997; Mathiowetz 1998). Likewise, respondents may provide answers in a range or other form that cannot be easily coded into the response categories. Third, respondents may provide one of two types of *nonsubstantive answers* by saying “don’t know” or refusing to answer altogether (Beatty and Herrmann 2002). These breakdowns in answering can occur because of the respondent, the interviewer, or characteristics of the questions themselves (Krosnick and Presser 2010; Schaeffer and Dykema 2011b).

Previous studies have demonstrated the joint effects of multiple question characteristics on question reliability and validity (e.g., Andrews 1984; Alwin 2007; Saris and Gallhofer 2007). Yet limited systematic attention has been given to examining the effects of multiple question, respondent, and interviewer characteristics across a full questionnaire on respondent answering behaviors (but see Holbrook, Cho, and Johnson 2006; Dykema, Schaeffer, Garbaski, Nordheim,

Banghart, et al. 2016; Holbrook, Johnson, Cho, Shavitt, Chavez, et al. 2016). Respondent behaviors during interviewer-administered surveys partially reveal mechanisms for creating reliable and valid answers. Thus, understanding the joint effects of multiple question and respondent characteristics on response behaviors is crucially important for understanding measurement errors.

This paper simultaneously examines the association of multiple question and respondent characteristics with six respondent behaviors in a telephone interview using cross-classified random effects logistic regression models. Because we have few interviewer characteristics available, we focus on question and respondent characteristics. We start by considering the cognitive response process, identifying question characteristics that are likely to affect each stage of this process. Next, we empirically evaluate the association between the question and respondent characteristics and the six behaviors. Finally, we discuss implications for questionnaire design and survey practice.

1.1 Respondent Behaviors and the Cognitive Response Process

To answer a survey question, respondents must comprehend the question, retrieve relevant information, make a judgment and map the answer to a given response category, and report the answer to the interviewer (Tourangeau, Rips, and Rasinski 2000). Figure 1 presents these four cognitive steps and potential behavioral reflections of these steps, adapted from models by Beatty and Herrmann (2002) and Ongena and Dijkstra (2007). For ease of visualization, the model is displayed in three panels, repeating the respondent behaviors.

Ideally, respondents move smoothly through the steps of the response process and provide an adequate (and accurate) substantive response, but sometimes breakdowns occur that lead to both biasing and variable measurement errors (Krosnick 1991; Tourangeau et al. 2000; Schaeffer and Dykema 2011b). In an interviewer-administered survey, respondents may or may not disclose these breakdowns to the interviewer. If they do disclose these breakdowns, they may do so at any stage by asking for clarification or by providing *nonsubstantive* (i.e., “don’t know,” “refuse”) or *problematic substantive responses* (qualified or uncodable answers).

Which behavior is most likely to occur depends on where the breakdown occurred in the response process. For example, several studies

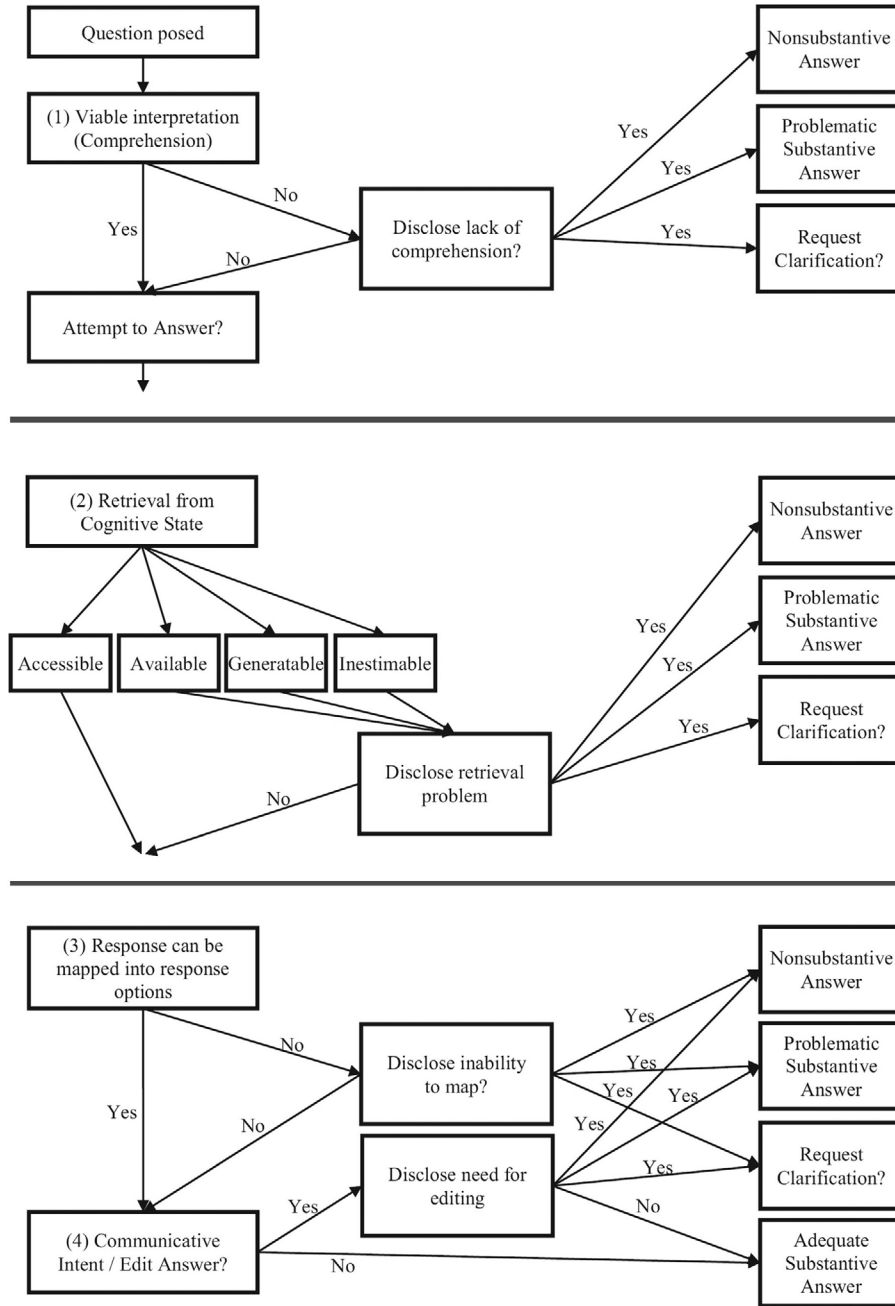


Figure 1. Flowchart Illustrating Cognitive Response Process and Respondent Behaviors (Adapted from Beatty and Herrmann 2002 and Ongena and Dijkstra 2007).

have shown that comprehension problems produce requests for clarification (Fowler 1992; Fowler and Cannell 1996; Holbrook et al. 2006). Likewise, problematic substantive answers have been linked to problems with both retrieval and mapping (Mathiowetz 1998; Holbrook

et al. 2006, 2016), and nonsubstantive answers have been linked to breakdowns at all stages of the cognitive response process (Beatty and Herrmann 2002). Thus, requests for clarification and nonsubstantive and problematic substantive answers indicate the risk of measurement errors (Hess, Singer, and Bushery 1999; Moore and Maynard 2002; Schaeffer and Maynard 2002; Ongena 2005; Holbrook et al. 2006; Ongena and Dijkstra 2007; Schaeffer and Dykema 2011a, 2011b). *Adequate answers* easily fit into the response categories and contain no observable signals that respondents are having difficulty, although may still be inaccurate or unreliable (Schaeffer and Dykema 2011b).

1.2 Question Characteristics That May Cause Respondent Answering Problems

Some question characteristics are particularly likely to cause comprehension breakdowns, while others may cause retrieval, judgment, or reporting breakdowns. Thus, we identify question characteristics by the stages of the response process we expect them to be most likely to affect, although some question characteristics may affect multiple stages of the response process and may be hard to disambiguate from other characteristics in a particular survey (e.g., attitude questions with ordinal response scales) (Dykema, Schaeffer, Garbarski, and Hout in press). Additionally, we expect that some question characteristics affect response behavior because they reflect more general learning about the questionnaire or respondent fatigue.

1.2.1 Comprehension. Comprehension difficulties may arise when respondents are asked to perform complex tasks or understand challenging vocabulary. The Flesch-Kincaid reading level is a commonly used measure of question comprehension difficulty and is associated with respondent behaviors indicating comprehension and mapping difficulties and with data quality problems (Holbrook et al. 2006, 2016; Velez and Ashworth 2007; Lenzner 2012, 2014; Olson and Smyth 2015). We expect that questions with higher reading levels will inhibit comprehension and, thus, be associated with more requests for clarification, fewer adequate answers, and higher rates of problematic substantive and nonsubstantive (especially “don’t know”) answers.

Longer questions or questions with transition statements require respondents to keep more information in their working memory,

burdening respondents (van der Zouwen 2000; van der Zouwen and Dijkstra 2002; Holbrook et al. 2006). On the other hand, these features may give respondents more time to think about the question (Cannell, Miller, and Oksenberg 1981). These counteracting forces result in mixed empirical evidence about question length and data quality. Empirically, introductory transition statements and longer questions decrease or have no clear effect on data quality (e.g., Knauper, Belli, Hill, and Herzog 1997; van der Zouwen 2000; van der Zouwen and Dijkstra 2002; Holbrook, Krosnick, Moore, and Tourangeau 2007; Saris and Gallhofer 2007). Question length is not consistently associated with respondent behaviors or other outcomes (Cannell et al. 1981; Alwin and Beattie 2016; Holbrook et al. 2016). As such, if length poses a burden, then we expect longer questions and those with transition statements to have higher rates of clarification requests, nonsubstantive responses, and problematic substantive response behaviors, and lower rates of adequate answers. If length or transition statements provide respondents with more time to think about a question, we expect the opposite associations.

If the question contains unknown terms, respondents are more likely to ask for clarification or respond with a “don’t know” response, although the effects of unknown terms on other respondent behaviors are mixed (Morton-Williams and Sykes 1984; Fowler 1992; Fowler and Cannell 1996; Johnson, O’Rourke, Chavez, Sudman, Warnecke, et al. 1996; Knauper et al. 1997; Holbrook et al. 2006; Olson and Smyth 2015). Adding definitions, either as part of the question text or as an optional statement read to the respondents, is one approach to dealing with unknown or technical terms. Simple definitions allow respondents time to think and may improve comprehension and the chances of ultimately providing adequate answers, although unclear or surprising definitions are more difficult to understand and result in “don’t know” responses (Conrad and Schober 2000; Tourangeau, Conrad, and Couper 2013) and longer response times (Olson and Smyth 2015). Therefore, we expect that unknown terms and definitions will increase the rate of clarification requests and the rate of problematic substantive and nonsubstantive answers and decrease adequate substantive answers.

1.2.2 Retrieval. The second step in the cognitive response process (box 2 in figure 1) is retrieval of information from memory. Information may be available (i.e., retrievable with little effort), accessible (i.e.,

retrievable with effort), generatable (i.e., can be formulated using related information in memory), or inestimable (i.e., not known and no information is available for generating an answer) (Beatty and Herrmann 2002, p. 73). While a question's content is likely the strongest determinant of information availability, question type— whether attitude, behavior, or demographic—may proxy for how readily information can be retrieved from memory.

We anticipate that the largest differences across question types will be between demographic and other types of questions. Demographic questions generally ask about readily available or accessible autobiographical facts and are therefore likely to pose few retrieval problems, increasing the likelihood of an adequate answer. Behavioral and attitudinal questions may require any of the retrieval types, depending on whether the question is unclear, has a long time frame, requires detailed information, requires respondents to construct an answer on the spot, or reconcile competing information on a particular domain (Tourangeau et al. 2000). As such, both behavioral and attitudinal questions are answered more slowly than demographic questions (Bassili and Fletcher 1991; Yan and Tourangeau 2008; Olson and Smyth 2015). The extra effort required for either attitudinal or behavioral questions should result in increased rates of problematic substantive and nonsubstantive answers relative to demographic questions (Fowler and Cannell 1996; Dykema et al. 1997; Ongena and Dijkstra 2007). Additionally, demographic questions are likely to be familiar and answered on previous surveys or forms, yielding fewer requests for clarification than attitudinal or behavioral items. Differences between behavioral questions and attitudinal questions depend on the specific content of these questions; we have no clear expectations for whether attitudinal or behavioral items are more difficult in this survey.

1.2.3 Judgment/Mapping. The next step of the response process is judgment (box 3 in Figure 1). In this step, respondents formulate an answer and map the answer to the response task. Judgment can break down if the response is determined to be inadequately certain or precise or if respondents have difficulty mapping it to the response options. We focus here on the mapping portion of judgment because we cannot parse out answer formulation with the available question characteristics.

Response option format is intrinsically linked to the difficulty of the mapping task. Given the narrative nature of open-ended text questions, any kind of response that corresponds to the question topic may be considered adequate. Therefore, we expect high rates of adequate answers for this question type. More problems can occur when mapping in open-ended numeric (interviewer types a number) or closed-ended (select one option) questions (Bradburn and Miles 1979; Blair and Burton 1987). In general, open-numeric, closed-nominal, and yes/no response option formats are all answered more quickly than open-ended text questions (Holbrook et al. 2007, 2016; Olson and Smyth 2015). However, open-ended numeric questions lead to more requests for clarification and mapping problems than other formats, likely because there is little up-front guidance about whether numeric or text responses are required (Holbrook et al. 2006). Closed-ordinal income questions can also be difficult because they ask respondents to map their answers into a very fine set of categories. Because of these difficulties, we expect more problematic substantive and nonsubstantive answers and more requests for clarification for open-numeric formats and closed-ended income questions than the other response option formats.

Questions with a greater number of response options take longer to answer, an indicator of difficulty (Yan and Tourangeau 2008; Olson and Smyth 2015), and affect question reliability and validity, although the magnitude and direction varies over studies (Andrews 1984; Saris and Gallhofer 2007, p. 241; Alwin, Baumgartner, and Beattie 2017). More response options require respondents to process the additional items, hold more in working memory, and do more fine-grained mapping into the categories. This may result in requests for clarification (e.g., “what were the choices again?”), problematic substantive answers, and “don’t know” answers.

When the response options do not match the question or concept asked about, respondents will have trouble judging how to answer the question (Houtkoop-Steenstra 2000; Dijkstra and Ongena 2006; Ongena and Dijkstra 2010; Olson and Smyth 2015). We expect a mismatch between the concept or task in the question and the task in the response options to yield higher rates of problematic substantive answers. Smyth and Olson (2016) found that experimentally mismatched question stems led to reduced rates of nonsubstantive responses in a telephone survey, possibly because respondents commit to answering

before discovering the mismatch (i.e., give a problematic substantive reflecting their initial understanding before discovering the mismatch). Thus, we expect mismatches between the question and response options to be associated with reduced rates of nonsubstantive answers. For the same reason, we also expect fewer requests for clarification on mismatched questions.

1.2.4 Reporting. The final step of the response process is reporting where respondents have to decide if they are willing to report their answer as is or edit it. A decision to report as is will result in an adequate response. If respondents feel the need to edit, most likely with sensitive questions, they will then need to determine whether or not to make the edit explicit by requesting clarification, implicit by providing a nonsubstantive or problematic substantive answer, or to not disclose the edit by providing an inaccurate adequate substantive answer. Respondents may be unwilling to accurately answer sensitive questions because the topic is intrusive or because of potential consequences of their answer (Tourangeau et al. 2000; Tourangeau and Yan 2007). Sensitive questions are answered more quickly than non-sensitive questions to quickly ease respondent burden, as well as appear appropriate to the interviewer (Tourangeau and Yan 2007; Olson and Smyth 2015; Fail, Schober, and Conrad 2016). Providing qualified answers or ranges for sensitive behaviors or asking for clarification may increase the sensitivity of these questions (e.g., admitting to engaging in a sensitive behavior such as having multiple sexual partners in a short period of time is bad enough, but suggesting that the number of partners one had during this short period is unknown may be even more detrimental to one's self-presentation). Because of this, we expect that respondents will be more likely to answer sensitive questions adequately to avoid potentially awkward interactions with the interviewer or to provide nonsubstantive answers to avoid revealing sensitive information and less likely to provide problematic substantive answers or ask for clarification. As such, an adequate response may not necessarily be accurate; it simply means there is no observable signal of a breakdown in the response process.

1.2.5 Survey fatigue and learning. Respondent answering behaviors may also be affected negatively from fatigue or positively from learning how to be good respondents as the survey proceeds. The position

of a question within a survey may be an indicator of either of these processes. As a survey continues, a respondent may experience fatigue that makes them less likely to optimize, resulting in satisficing (Narayan and Krosnick 1996; Galesic and Bosnjak 2009), which can take the form of increased problematic substantive and nonsubstantive answers and fewer requests for clarification on later items. On the other hand, respondent experience with the questionnaire may provide them with training on how to answer questions. In this case, later questions may also yield fewer requests for clarification (as respondents need less help) but fewer problematic substantive answers and more adequate answers. Empirical evidence on placement of items on response timing and other data quality outcomes is mixed and depends on mode (e.g., Andrews 1984; Holbrook et al. 2007; Saris and Gallhofer 2007; Yan and Tourangeau 2008; Olson and Smyth 2015). Holbrook et al. (2016) found that mapping problems decrease as the questionnaire continues, lending evidence to a learning hypothesis. We anticipate that the rate of requests for clarification and problematic substantive and nonsubstantive answers will decrease for later questions and that the rate of adequate substantive answers will increase.

Battery questions are a set of items connected by a single introduction and shared response options. In general, battery items are less reliable than nonbattery items, and the first item in a battery differs in its measurement error properties from later items in the battery (Alwin 2007; Saris and Gallhofer 2007; Schaeffer et al. 2015; Alwin and Beattie 2016). As with question location overall, in battery items, respondents may become trained about how to answer as they go along. Alternatively, respondents may forget the response options or the question prompt later in a battery, leading to more problematic substantive behaviors. Therefore, under the learning hypothesis, we expect the first question in a battery to have more problematic substantive and nonsubstantive answers and more requests for clarification than later questions in the battery, but we expect the opposite to occur under the fatigue hypothesis.

1.3 Respondent Characteristics

Characteristics of the respondents may affect all parts of the cognitive response process. Education level and age of respondent are common indicators of cognitive abilities (Krosnick 1991; Narayan and Krosnick

1996; Knauper et al. 1997; Holbrook et al. 2007). Those with low cognitive ability (less education and older) are more likely to experience breakdowns in the response process and need to request clarification or report “don’t know” answers, possibly because it is more difficult for them to understand the questions (Knauper et al. 1997, but see Holbrook et al. 2006). We anticipate higher rates of requests for clarification, problematic substantive answers, and nonsubstantive answers for older and less educated respondents.

In addition, people who are distracted while being interviewed will likely respond with inadequate answers. Distractions are commonly posited as a problem for telephone interviews (Schwarz, Strack, Hippler, and Bishop 1991; Lynn and Kaminska 2013). We include the number of people in the household as an indicator of the risk of distractions, anticipating more breakdowns and more requests for clarification, problematic substantive answers, and nonsubstantive answers for individuals in larger (more distracted) households.

We now turn to empirically examining whether and how these theoretically motivated question and respondent characteristics are associated with telephone survey respondent behaviors.

2. Methods

The data for this paper come from the Work and Leisure Today (WLT) study (AAPOR RR3 = 6.3 percent, AAPOR 2016). The WLT study, conducted by AbtSRBI, interviewed adults in the United States with landline telephone numbers during the summer of 2013, omitting thirty-eight percent of adults who lived in cell phone-only households (Blumberg and Luke 2010).

Interviews were digitally recorded and transcribed. The 449 transcribed interviews were behavior coded at the conversational turn level using Sequence Viewer software (Dijkstra 1999). Behavior coding is an objective, reliable method used to identify problems with specific questions in a survey (Belli and Lepkowski 1996; Fowler and Cannell 1996; Maynard and Schaeffer 2002; Fowler 2011).

Behavior coding was conducted by trained undergraduates, with two master coders independently coding a ten percent subsample. We use three coded attributes for this paper, including the speaker (interviewer or respondent, kappa = 0.998) and, if the respondent was

speaking, whether he or she provided an answer, asked for clarification, or gave feedback ($\kappa = 0.89$). When an answer was provided, coders determined if it was adequate (i.e., “codable” or met the response task), qualified (e.g., “about 5”), uncodable (e.g., ranges, inappropriate responses), a “don’t know” or a refusal (coded separately, hereafter DK/REF when combined) ($\kappa = 0.78$).¹ Examples of the behaviors coded into each category are presented in Table 1.

We examine six dichotomous dependent variables representing the respondent behaviors on their first conversational turn immediately after the interviewer asked the question: adequate answer, two types of problematic substantive answers (qualified and uncodable), non-substantive answers (DK or refusal combined and “don’t know” on its own), or a request for any type of clarification. Respondents may have more than one type of answering behavior on a given question; examining only the first turn avoids contamination of the answering behavior by any other interaction with the interviewer. A total of 20,936 first respondent conversational turns were coded; forty-nine (0.23 percent) are excluded because of unintelligible audio, leaving $n = 20,887$ turns. “Don’t know” and refusal responses each occurred on less than one percent of respondents’ first conversational turns (DK = 0.84 percent; refuse = 0.82 percent); we combine these two non-substantive responses into an overall DK/ REF response and also report the DK model on its own.² Overall, 67.9 percent of respondent first turns had an adequate answer (adequate answer = 1, all other behaviors = 0), 6.2 percent were qualified (qualified answer = 1, all other behaviors = 0), 11.5 percent were uncodable answers, 1.7 percent were DK/ REF, 0.84 percent were “don’t know” alone, and 8.6 percent were requests for clarification. The remaining 4.1 percent of first conversational turns were one of fifteen other types of behavior (e.g., personal disclosures).

1. Interruptions were coded as a separate field; interviewers could have interrupted respondents on any of the respondent behaviors. Approximately three percent of adequate answers, six percent of qualified answers, twelve percent of uncodable answers, eight percent of DK/REF answers, and seven percent of respondent requests for clarification were interrupted.
2. Most of the items in the survey did not have any initial refusals. Thus, the model predicting REF did not converge due to sparse cells. See appendix figure A.1.

Table 1. Examples of Respondent Statements for Respondent Behaviors

<i>Respondent behavior</i>	<i>Example statements</i>
Adequate	<ul style="list-style-type: none"> • 7 days a week. • Oh, it, I do every other day, so it was, like, four, four days a week. • None, cause I don't drive. • None. • Zero.
Qualified	<ul style="list-style-type: none"> • Probably, uh, 2 days a week. • I would say about 6. • Sigh. Maybe eight. • Mmm, oh geez um, maybe five hours. • Uh not that much, uh probably, probably 10 maybe. No more than that.
Uncodable	<ul style="list-style-type: none"> • Like, thousands, literally thousands of miles. (Q10) • Well I worked hard. I figure that's exercise I never exercise seriously I mean going any place to do it. (Q10) • Tsk uh, I, I would, I would, I have never hunted, so the only thing I've done is fish, so, uh, I could put, I would say a 1 or 2. (Q13D) • I quit smoking when I was 25. (Q21B)
DK/REF	<ul style="list-style-type: none"> • Uh I really don't know uh . . . • Oh I have no idea. • I don't know what that is. • Uh, I'll skip that question. • I'm not gonna tell you that. • That's personal. Nobody's business. • I don't want to answer that.
Clarification requests	<ul style="list-style-type: none"> • Uh, I would say, probably . . . Oh, I work some . . . Now, you're asking about a week? • Uh, how many hours? • What is the question again? • Okay, what do you define as leisure time? • What's that? • Is it 1 to 5 or . . . ?

Q10 asked about number of days of exercise during the last week; Q13D asked about how much the respondent enjoyed fishing or hunting on a scale from 1 to 5, and Q21B asked how many times the respondent smoked a cigarette during the past seven days.

Our first set of independent variables is question characteristics (see Table 2 and supplementary materials). These characteristics were coded by two independent graduate student coders (kappa range from 0.85 to 1.00) with discrepant codes resolved by two of the authors (Olson and Smyth 2015). Question characteristics potentially causing comprehension problems include question reading level, question length, whether there is a transition statement, unknown terms, and definitions in the question stem. Whether the question is

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

	<i>n</i>	<i>Mean/%</i>	<i>SD</i>
Question characteristics			
Factors affecting comprehension			
Question length	54	14.56	12.71
Question reading level	54	6.64	4.76
Transition statement in stem	54	13.0%	
Unknown terms in question	54	3.7%	
Definitions in stem	54	18.5%	
Factors affecting retrieval			
Question type			
Attitude (all closed-ordinal)	17	31.5%	
Behavior	23	42.6%	
Demographic	14	25.9%	
Factors affecting judgment			
Response option format			
Open-ended text	5	9.3%	
Open-ended numeric	17	31.5%	
Closed-nominal	6	11.1%	
Closed-ordinal (attitudinal)	17	31.5%	
Closed-ordinal (income only)	1	1.8%	
Yes/no	8	14.8%	
Number of response options	54	3.39	3.49
Mismatch between question and response options	54	13.0%	
Factors affecting reporting			
Sensitivity	54	13.0%	
Fatigue versus learning			
Battery position			
1 st in battery	4	7.4%	
Later in battery	18	33.3%	
Not in battery	32	59.3%	
Question sequence	54	23.22	14.72
Respondent characteristics			
Respondent age	449	61.34	16.72
Education = High school degree or less	449	29.2%	
Number of people in respondent's household	449	2.17	1.34
Respondent controls			
Female	449	63.9%	
Employed	449	41.0%	
Use internet	449	68.8%	
Interviewer controls			
Female	22	54.6%	
White	22	40.9%	
Employed 1+ year	22	68.2%	

an attitude, behavior, or demographic question is the proxy for potential retrieval problems.

Judgment/mapping may be affected by the response option format, the number of response options, or a mismatch between the stem and the response options. One limitation of this survey is that all of the

attitudinal questions used a closed-ended ordinal format. The only non-attitudinal ordinal question was the income question, which asked respondents to stop the interviewer during the reading of a long list of response options. Thus, we cannot fully disentangle the effects of attitudinal question type and closed-ended ordinal response option format. As a result, we discuss their simultaneous effects here as question type under retrieval (in both the text and results tables), comparing them with behavioral and demographic questions (but noting the limitation). We include income as the only closed-ended ordinal question in our discussion of response option formats.

Question sensitivity is likely to affect the editing/reporting stage. The sensitive questions in this survey asked about whether the respondent had ever been fired from a job, the number of times during the last seven days that the respondent drank alcohol, had sex, and looked at adult websites, the number of parking and speeding tickets received during the last year, and income. We also include indicators for whether the question is the first or a later question in a battery and question sequence (i.e., first question answered assigned 1, second question answered assigned 2, etc.) as possible indicators of fatigue or learning.

Respondent characteristics of age, measured as a continuous variable, and education level, measured as high school degree or less ($= 1$) versus some college or more ($= 0$), serve as proxies for cognitive difficulty. The number of people in the respondent's household is included as a proxy for distractions.

In addition to these key independent variables, we also control for several respondent ($n = 449$) and interviewer ($n = 22$) characteristics. We control for respondent sex (female $= 1$) to account for potential conversational differences between males and females (Goldshmidt and Weller 2000), and employment status (employed $= 1$) and internet status (uses internet $= 1$) to account for skip patterns in the questionnaire. Interviewer sex (female $= 1$), race (white $= 1$), and experience (1)year of experience $= 1$) are also included as control variables.

We use cross-classified random effects logistic regression models (Raudenbush and Bryk 2002; Beretvas 2011) to simultaneously evaluate the association of multiple question and respondent characteristics with respondent behaviors. Each behavior is cross-classified by respondents and by questions, with questions and respondents nested within interviewers.

We adapt notation by Beretvas for a three-level cross-classified linear model (2011, pp. 330–331) to a cross-classified logistic regression model. The base model predicts the logit of the probability of a particular respondent behavior occurring on each question, where $Y_{i(j_1, j_2)k} = 1$ indicates that the behavior occurs, as a function of an overall mean (γ_{0000}) plus random effects due to the respondent ($u_{0j_1 0k}$), the question ($u_{00j_2 k}$), and the interviewer (u_{000k}). We assume that the random effects are normally distributed with mean zero and variance τ_{uj_1} , τ_{uj_2} , and τ_{uk} , respectively (Beretvas 2011, p. 330):

$$\text{logit}(\Pr(Y_{i(j_1, j_2)k} = 1)) = \gamma_{0000} + u_{000k} + u_{0j_1 0k} + u_{00j_2 k}$$

We calculate the proportion of the variance in $\text{logit}(\Pr(Y_{i(j_1, j_2)k} = 1))$ associated with questions, respondents, and interviewers. For example, we use:

$$\rho_{resp} = \frac{\hat{\tau}_{uj_1}}{\hat{\tau}_{uj_1} + \hat{\tau}_{uj_2} + \hat{\tau}_{uk} + \pi^2/3}$$

for the proportion of variance due to respondents; we modify this equation for the variance due to interviewers and questions.

Question, respondent, and interviewer characteristics are then added to the base model for a final model for each behavior:

$$\begin{aligned} \text{logit}(\Pr(Y_{i(j_1, j_2)k} = 1)) = & \gamma_{0000} \\ & + \sum_{s=1}^q \beta_s \text{Question_char}_{j_2} \\ & + \sum_{m=1}^p \beta_m \text{Respondent_char}_{j_1} \\ & + \sum_{t=1}^r \beta_t \text{Iwer_char}_k \\ & + u_{000k} + u_{0j_1 0k} + u_{00j_2 k} \end{aligned}$$

All of the models are estimated using restricted maximum likelihood estimation in Stata 15.0 *xtmelogit* with random intercepts for questions, respondents, and interviewers (Rabe-Hesketh and Skrondal 2012). All continuous predictors are grand-mean centered.

Table 3. Null Models of Respondent Behaviors on First Respondent Conversational Turns

	<i>Interviewer</i>	<i>Question</i>	<i>Respondent</i>	<i>LR test</i>
Adequate				
Variance Component (SD)	0.223	1.371	0.763	6101.72****
Proportion of Total Variance	0.009	0.324	0.100	
Qualified				
Variance Component (SD)	0.000	1.084	0.854	1393.21****
Proportion of Total Variance	0.000	0.226	0.140	
Uncodable				
Variance Component (SD)	0.155	1.127	0.847	2219.06****
Proportion of Total Variance	0.005	0.240	0.135	
DK/REF				
Variance Component (SD)	0.397	1.488	0.803	685.63****
Proportion of Total Variance	0.025	0.351	0.102	
Don't know				
Variance Component (SD)	0.325	1.580	0.783	313.85****
Proportion of Total Variance	0.032	0.759	0.186	
Request for Clarification				
Variance Component (SD)	0.235	1.336	0.565	2891.46****
Proportion of Total Variance	0.010	0.328	0.059	

n = 20,887 first conversational turns; **** *p* < 0.0001

3. Results

Table 3 shows the proportion of variance for the null models for each respondent behavior. The proportion of variance for the interviewer is virtually zero in all of the models, indicating little variability across interviewers in how respondents answer survey questions. However, we see significant variation across questions in each of these response behaviors. Between 22.6 percent and 75.9 percent of the total variance in these response behaviors is due to the questions. In virtually every model, this is at least twice as large as the proportion of total variance due to respondents (5.9 percent to 18.6 percent). Thus, questions contribute more variability to respondent behaviors than respondents themselves.

3.1 Question Characteristics

Given the large number of question characteristics and response behaviors, we summarize our original predictions and the results from the models in Table 4. The full models are presented in Appendix Table A.3.

Table 4. Predicted Associations and Empirical Associations for Three Groups of Response Behaviors

	Prediction				Results				
	Adequate	Problematic substantive		Clarification	Adequate	Problematic substantive		Clarification	
		Qualified	Uncodable			Qualified	Uncodable		
Question characteristics									
Comprehension									
Question reading level	-	+	+	+	-	n/s	+	n/s	+
Question length	+/-	+/-	+/-	+/-	n/s	n/s	n/s	n/s	n/s
Transition statement in stem	+/-	+/-	+/-	+/-	n/s	n/s	n/s	n/s	+
Unknown terms in question	-	+	+	+	-	-	-	+	+
Definitions in stem	-	+	+	+	n/s	n/s	+	n/s	n/s
Retrieval									
Question type									
Demographic	+	-	-	-	+	-	n/s	-	n/s
Attitude (closed-ordinal)	+/-	+/-	+/-	+/-	n/s	n/s	n/s	n/s	n/s
Behavior	+/-	+/-	+/-	+/-					
Mapping									
RO format									
Open-ended text	+	-	-	-					
Open-ended numeric	-	+	+	+	-	+	+	n/s	n/s
Closed-ordinal (income only)	-	+	+	+	-	+	+	+	n/s
Closed-nominal	+	-	-	-	n/s	n/s	n/s	n/s	-
Number of ROs	-	+	+	+	n/s	n/s	+	n/s	n/s
Mismatch between question and ROs	-	+	+	-	+	n/s	n/s	n/s	-
Reporting									
Sensitivity	+	-	-	-	+	-	-	-	-
Fatigue versus learning									
Battery position									
1st in battery	+/-	+/-	+/-	+/-	n/s	-	n/s	n/s	n/s
Later in battery	+/-	+/-	+/-	+/-	n/s	n/s	+	-	n/s
Question position	+	-	-	-	n/s	n/s	n/s	+	n/s
Respondent characteristics									
Respondent age	-	+	+	+	-	n/s	+	n/s	+
Education = HS degree or less	-	+	+	+	-	n/s	+	n/s	n/s
# of people in respondent's HH	-	+	+	+	n/s	n/s	n/s	n/s	n/s

+/- indicates competing hypotheses for a variable or no clear association predicted.

We anticipated that most of the factors affecting comprehension would increase the probability of nonsubstantive and problematic substantive answers and requests for clarification and decrease the probability of adequate answers. As expected, questions with higher reading levels are less likely to yield an adequate answer and more likely to yield a request for clarification and uncodable answers, but they are unexpectedly not associated with nonsubstantive answers. Question length and transition statements are not associated with any of the substantive or nonsubstantive response behaviors, likely reflecting the competing mechanisms of burden and time to think about an answer; transition statements do increase requests for clarification. Consistent with our predictions, questions with unknown terms decrease adequate substantive answers and problematic substantive answers, and increase nonsubstantive answers and requests for clarification. Contrary to expectations, questions with definitions in the question stem increase the probability of uncodable answers, but have no association with the other behaviors. In all, unknown terms are consistently related to response behaviors, with other comprehension measures primarily increasing rates of clarification requests and uncodable answers.

Question type—our indicator of the type of retrieval task—is associated with qualified and “don’t know” answers. As expected, demographic questions are less likely to have qualified or “don’t know” answers than behavioral questions. There is no difference in response behaviors between the attitudinal items (which all have closed-ordinal response options) and behavioral or demographic items in this survey (Appendix Table A.4).

Next, as expected, each of the proxies for mapping problems is associated with the response behaviors (overall tests and pairwise comparisons in appendix table A.4). As predicted, questions with open-ended numeric response options had fewer adequate answers and more qualified and uncodable responses relative to open-ended text questions. This is likely because open-ended text questions have fewer restrictions on what constitutes an adequate answer, and qualifying information in the answer is acceptable. Similarly, the income question with closed-ordinal response options had fewer adequate answers and higher rates of problematic substantive and nonsubstantive behaviors than other types of questions. This is not terribly surprising;

the income question had a very long list of categories and required respondents to interrupt the interviewer when she read the appropriate category, a difficult task. Interestingly, closed-nominal response options were similar on all of the response behaviors to open-ended text questions, except that they had fewer clarification requests. In this survey, closed-nominal questions tended to ask about known and readily available information. Yes/no questions had lower rates of qualified and “don’t know” answers than open-ended text responses.

Also, as expected, questions with more response options have higher rates of uncodable answers. This may be a result of difficulty finding the “right” response option for the task at hand. There is no association between the number of response options and the other behaviors.

Somewhat surprisingly, a mismatch between the stem and the response options increases the probability of adequate answers, but decreases the probability of clarification requests. It is possible that interviewers preemptively modified the question and read the response options or clarified the concept needed before the respondent could attempt an answer to attempt to address the mismatch.

Next, we examine the proxy for difficulties with the reporting step. The sensitivity of a question is significantly associated with all of the respondent behaviors. As expected, sensitive questions had higher rates of adequate answers, but lower rates of qualified, uncodable, and “don’t know” answers and clarification requests. When respondents are uncomfortable answering, the quickest and least intrusive strategy may be to give an adequate answer; all other types of answers are likely to trigger interviewer probes and prolong the uncomfortable interaction and may also reflect poorly on the respondent.

We see some evidence of fatigue and learning effects. Later questions in a battery have higher rates of uncodable answers (consistent with fatigue), but lower rates of DK answers (consistent with learning). Thus, respondents may have learned (incorrectly) how to answer these questions from the first question in the battery. Question position is associated only with nonsubstantive answers.

3.2 Respondent Characteristics

As expected, older respondents were less likely to provide adequate answers and more likely to provide uncodable answers and request clarification. Also as anticipated, respondents with high school education

Table 5. Percent Reduction in Variance for Interviewers, Questions, and Respondents

	<i>Null model (SD)</i>	<i>Full model (SD)</i>	<i>Diff in variance (Null²–Full²)</i>	<i>% reduction in variance (Diff in variance/Null²)</i>
Adequate				
Interviewer	0.223	0.175	0.019	38%
Question	1.371	0.686	1.409	75%
Respondent	0.763	0.662	0.144	25%
Qualified				
Interviewer	0.000	0.000	0.000	n/a
Question	1.084	0.542	0.881	75%
Respondent	0.854	0.821	0.055	8%
Uncodable				
Interviewer	0.155	0.139	0.005	20%
Question	1.127	0.601	0.909	72%
Respondent	0.847	0.634	0.315	44%
Don't know				
Interviewer	0.325	0.268	0.034	32%
Question	1.580	0.649	2.075	83%
Respondent	0.783	0.777	0.009	2%
DK/REF				
Interviewer	0.397	0.275	0.082	52%
Question	1.488	0.997	1.220	55%
Respondent	0.803	0.787	0.025	4%
Clarification request				
Interviewer	0.235	0.137	0.036	66%
Question	1.336	0.715	1.274	71%
Respondent	1.565	0.561	2.135	87%

or less are less likely to provide an adequate answer and more likely to provide problematic substantive answers. The number of people in a respondent’s household was not associated with any of the respondent behaviors, failing to support the distraction hypothesis.

3.3 Explained Variance Components

Table 5 shows the percent of variance explained at the question, respondent, and interviewer level by the models shown. Across the behaviors, the covariates explained fifty-five percent or more of the variation due to the questions. This is a sizable effect—simply knowing the modest package of question characteristics examined here (as well as a few respondent and interviewer characteristics) explains between about half and over three-quarters of the variation that we can expect in respondent answering behaviors and clarification requests across questions.

The models are less successful at explaining variability across respondents in their response behavior. The covariates explain between two percent and eighty-seven percent of the variation across the behaviors. There is still more to learn about how different characteristics of respondents contribute to these behaviors. Between twenty percent (uncodable answers) and about sixty-six percent (clarification requests) of the interviewer-level variance was explained by these covariates.

4. Discussion

This analysis aims to answer a simple question: are question and respondent characteristics associated with initial respondent behaviors after the interviewer reads a survey question? These behaviors are important because they are indicators of breakdowns in the cognitive response process and thus provide insights into risks of measurement error in reports. The answer is yes—question and respondent characteristics are associated with respondent behaviors. We are able to explain a substantial proportion of the variability of these behaviors across both questions and interviewers, but have less success in explaining the variability across respondents. There are four main findings.

First, the survey research literature focuses largely on improving comprehension of questions to minimize the risk of measurement error (Fowler 1992, Dillman, Smyth, and Christian 2014). Here, measures of question characteristics that are likely to cause comprehension difficulties were associated with requests for clarification and problematic substantive answers. In particular, although question length was not associated with any of the response behaviors, our results suggest that survey designers are well-advised to avoid unknown terms in questions, to write questions with lower reading levels, and to use definitions and transition statements only when necessary to minimize problematic response behaviors and requests for clarification.

Second, although we see weak evidence of differences in respondent behaviors across question types (our measure of potential retrieval problems), one of our measures for mapping difficulties (the type of response options) was related to all of the respondent behaviors. In this survey, as expected, the initial response task was much more

straightforward for demographic items. Additionally, open-ended numeric questions and the closed-ended income question posed the greatest difficulties for respondents, and closed-ended nominal, yes/no, and open-ended narrative questions posed the fewest difficulties. Unfortunately, all of our attitudinal items had closed-ended ordinal response options. Although this is typical of attitudinal items, we cannot disentangle whether other types of response tasks (e.g., yes/no, semantic differential) would have similar outcomes. Future research should investigate this in questionnaires containing varying response tasks for attitudinal items.

Surprisingly, questions where there was a response task or concept mismatch between questions and response options had a higher rate of adequate answers and fewer requests for clarification. From this analysis, we do not know whether interviewers preemptively changed the question wording in order to address these task and conceptual misalignments (similar to an interviewer decision to read parentheticals; Dykema et al. 2016). Future research will examine the interviewer-respondent interaction on mismatched questions in more detail.

Third, our findings are most counter to conventional wisdom for sensitive questions. As we expected, they led to fewer requests for clarification than nonsensitive questions, but they were also less likely to manifest problematic answering behaviors. This could be because some respondents have not experienced the sensitive behavior or their levels of experience are within socially acceptable levels, making it easy for them to register an adequate answer. Others may be editing their response behaviors (and possibly their answer) to appear appropriate to the interviewer (Tourangeau and Yan 2007). These respondents may provide an erroneous adequate answer that falls within socially acceptable levels for the behavior in question. Under both of these hypotheses—truly not experiencing the behavior and editing responses—we would expect that answering behaviors are associated with the answers that people give. Alternatively, some respondents may determine their true answer is less embarrassing or problematic than implying that they are unsure whether or how often they have engaged in the behavior or than requesting clarification about the meaning of a sensitive behavior. For instance, the average number of drinks during the last week for respondents who provided an initial adequate response is 0.55, compared to 1.65 drinks for those who did not provide an initial adequate response, but eventually provided an answer (although about thirty-three percent of this group never

provided an answer to this question). These findings are especially important because sensitive questions are often examined in isolation in a mode comparison (Tourangeau and Smith 1996) or with limited contrast to other nonsensitive questions. Future research should delve into other types of interactional behaviors such as laughter or disfluencies that may provide a cue to problematic answering and how the final answers provided by the respondents are associated with their response behaviors.

Fourth, we found evidence of both respondent fatigue and learning hypotheses. Nonsubstantive responses increased for later items in the questionnaire, supporting a fatigue hypothesis. On battery items, items other than the first in the battery have higher rates of uncodable answers (supporting fatigue) and lower rates of “don’t know” responses (supporting learning).

Our findings lend support to common recommendations for questionnaire design, such as using easier question reading levels and avoiding unknown terms. Unfortunately, survey designers have limited control over many of the question characteristics that turned out to be significantly associated with respondent answering behaviors. For example, the type of response options is often dictated, to a large extent, by the research goals. Somewhat disappointingly, our findings suggest that beyond these few strategies, little more can be done by way of questionnaire design to improve respondent answering behaviors.

Instead, our findings suggest that the most fruitful avenue is to improve interviewer training on how to address these respondent problems when they are disclosed to the interviewer. In particular, interviewers should be trained to anticipate different kinds of respondent problems on different types of questions, such as initial qualified or uncodable answers on open-ended numeric or income questions. Survey designers could help interviewers by including instructions on the screen about appropriate probes or other clarification and verification behaviors. How interviewers resolve these problems is not examined here. Future research should examine the unfolding of the interviewer–respondent interaction after these initial responses are given to better understand what types of interviewer interventions can successfully transition these problematic responses to adequate answers.

This study has limitations. Excluding cell phone respondents (resulting in an older respondent pool) likely resulted in higher rates of

problematic respondent answering behaviors; future research should include both landline and cell phone samples. Nevertheless, we have little reason to believe that cell phone respondents would behave much differently to these same question stimuli, given no clear indication of measurement error differences between landline and cell phone respondents (AAPOR 2010). This is also a survey on one topic with one set of questions. Future research should examine surveys that have different question topics. Finally, this particular behavior coding scheme provides an important look across a wide variety of question types and topics, but it necessarily omits item- or question-type specific response behaviors that may be particularly revealing about how respondents answer particular types of questions. Future work will take a more in-depth, qualitative look at how these breakdowns manifest for individual types of survey questions.

This study also was limited to examining the question characteristics that were present in this particular 15-minute long telephone survey. For example, questions with show cards were necessarily omitted. Therefore, we cannot make inference to those types of questions or question characteristics. Additionally, question characteristics appear as a package (Dykema et al. in press), and as such, we cannot disentangle how different question features interact or moderate each other (e.g., attitudinal questions with different types of response options). Despite these limitations, much of the work on respondent-interviewer interaction is conducted in face-to-face surveys and looks at one question at a time (e.g., Suchman and Jordan 1990; Dykema et al. 1997), and less is known about telephone surveys, especially those conducted in a contemporary context (but see Dykema et al. 2016). While this study cannot assess some question types and question characteristics, it extends our knowledge about respondent-interviewer interaction to contemporary telephone surveys across multiple items. This observational study allows new insights that cannot emerge when question features are examined in isolation through experimental designs alone (for example, sensitive questions having fewer interactional problems than nonsensitive questions). Future research should take advantage of the strengths of both of these methods by embedding questionnaire design experiments in studies with coding of the respondent-interviewer interaction and looking across all survey questions simultaneously. Doing so will allow researchers to obtain a pure measure of how question features affect the interaction

between interviewers and respondents, as well as how these features act relative to other questions in a questionnaire, with the goal of improving survey data quality.

Acknowledgments — This work was supported by the National Science Foundation [SES-1132015 to Kristen Olson]. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank the Bureau of Sociological Research at the University of Nebraska-Lincoln for transcription and two anonymous reviewers and the editor for comments on a previous draft.

Appendix

Table A.1. Eligible Codes for Respondent Initial Actions

<i>Short description</i>	<i>Definition</i>
Answer provided	The respondent provides an answer
Clarification	The respondent requests clarification or definition
Feedback	The respondent provides feedback or other response (e.g., digression)

Table A.2. Eligible Codes for Assessments of the Respondent Answering Behaviors

<i>Short description</i>	<i>Definition</i>
Adequate answer	The respondent provides an adequate answer to the question. This is an answer that answers the question and can be coded into the response categories or response format in the questionnaire.
Qualified answer	The respondent provides an answer with a qualifier that shows uncertainty. This could be "about," "approximately," "I guess," "maybe," "kind of," "I believe," "basically," "not that I know of," "around" etc.
Uncodable answer	The respondent provides an answer that cannot be coded into the response categories.
Don't know	The respondent states that they do not know or don't remember the answer.
Refuse	The respondent refuses to answer the question.
Answers previous question	The respondent says that they have an answer to a previous question or continues to answer a previous question after the interviewer has moved to another question.
Agrees with interviewer	The respondent agrees with the interviewer, either at verification or as a method of showing attention or understanding. This is also used when the respondent provides an adequate answer to an interviewer's clarification or verification. For example, the interviewer asks "Is that a 5?" and the respondent says, "Yes."
Disagrees with interviewer	The respondent disagrees with the interviewer. This is also used when the respondent provides an adequate answer to an interviewer's clarification or verification but does not agree with the interviewer's clarification or verification. For example, the interviewer asks "Is that a 5?" and the respondent says, "No."

Table A.3. Full Model Coefficients and Standard Errors (in Parenthesis) Predicting Respondent Behaviors on First Respondent Conversational Turn

	Adequate		Problematic substantive		Nonsubstantive		
			Uncodable		DK/REF	Don't know	Any clarification request
	Qualified	Unqualified	DK/REF	Don't know	DK/REF	Don't know	Any clarification request
Comprehension							
Question reading level	-0.050+ (0.029)	0.032 (0.026)	0.047+ (0.027)	-0.034 (0.051)	0.021 (0.049)	0.057+ (0.033)	
Question length	-0.003 (0.013)	-0.014 (0.011)	0.015 (0.012)	-0.017 (0.025)	-0.038 (0.024)	0.006 (0.014)	
Transition statement in stem	-0.407 (0.383)	-0.055 (0.349)	-0.258 (0.356)	-0.198 (0.718)	0.817 (0.665)	0.699+ (0.419)	
Unknown terms in question	-1.855* (0.750)	-2.064** (0.670)	-2.913**** (0.688)	3.609** (1.239)	2.551* (1.081)	2.656** (0.807)	
Definitions in stem	-0.568 (0.389)	0.273 (0.330)	0.946** (0.352)	-0.692 (0.706)	-0.055 (0.642)	0.313 (0.428)	
Retrieval							
Type of question (behavior ref.)	0.657+ (0.394)	-1.169** (0.355)	-0.220 (0.372)	-1.122 (0.757)	-2.115** (0.784)	-0.024 (0.444)	
Demographic	0.199 (0.528)	-0.409 (0.459)	0.092 (0.489)	0.206 (0.925)	0.029 (0.839)	-0.479 (0.581)	
Judgment/Mapping							
RO format (open-ended ref.)	-1.501**** (0.412)	0.817* (0.352)	1.491**** (0.380)	0.604 (0.690)	0.150 (0.627)	0.503 (0.450)	
open-ended numeric	-3.619**** (1.014)	3.360**** (0.861)	1.569+ (0.921)	4.536** (1.680)	5.433**** (1.548)	0.685 (1.116)	
closed-ordinal (income only)	0.134 (0.504)	-0.682 (0.468)	0.282 (0.469)	-0.710 (0.945)	-14.472 (368.408)	-1.422* (0.590)	
closed-nominal	0.688 (0.447)	-0.882* (0.398)	-0.212 (0.418)	-0.948 (0.821)	-1.559+ (0.817)	-0.795 (0.499)	
yes/no	-0.055 (0.036)	0.034 (0.030)	0.090** (0.033)	-0.004 (0.058)	-0.038 (0.051)	0.016 (0.041)	
Number of ROs	0.914** (0.331)	-0.443 (0.287)	-0.123 (0.301)	0.014 (0.561)	-0.016 (0.569)	-1.301** (0.376)	
Mismatch between question and ROs							

Table A.3. Full Model Coefficients and Standard Errors (in Parenthesis) Predicting Respondent Behaviors on First Respondent Conversational Turn (continued)

Clarification	Adequate		Problematic substantive		Nonsubstantive		
	Qualified	Uncodable	DK/REF	Don't know	Any clarification request		
Reporting							
Sensitivity	1.757*** (0.416)	-1.799*** (0.369)	-1.282*** (0.386)	-0.787 (0.768)	-1.243+ (0.698)	-0.831+ (0.459)	
Fatigue versus Learning							
Battery items							
1st question in battery	-0.052 (0.491)	-0.758+ (0.426)	0.536 (0.443)	-0.799 (0.910)	-0.975 (0.731)	0.421 (0.525)	
Later question in battery	-0.221 (0.412)	-0.080 (0.359)	1.016** (0.379)	-0.943 (0.726)	-1.560* (0.367)	0.297 (0.449)	
Question sequence number	0.011 (0.009)	-0.012 (0.009)	-0.003 (0.009)	0.036+ (0.021)	0.020 (0.023)	-0.013 (0.011)	
Respondent Characteristics							
R's age	-0.011*** (0.003)	0.004 (0.004)	0.013*** (0.003)	0.005 (0.006)	0.005 (0.007)	0.006+ (0.003)	
R's education (hs or less = 1)	-0.238** (0.086)	0.193 (0.119)	0.271** (0.090)	0.081 (0.172)	0.319 (0.205)	-0.045 (0.094)	
Number of people in R's household	0.051 (0.032)	-0.027 (0.045)	-0.045 (0.035)	-0.067 (0.064)	-0.025 (0.078)	-0.006 (0.035)	
Respondent Controls							
R's sex (female = 1)	0.101 (0.077)	-0.213* (0.108)	-0.064 (0.083)	0.201 (0.155)	0.342+ (0.198)	0.126 (0.084)	
R's employment status (employed = 1)	0.230* (0.104)	-0.042 (0.135)	-0.573*** (0.115)	0.220 (0.220)	0.374 (0.260)	0.172 (0.120)	
R's internet status (use internet = 1)	0.084 (0.094)	0.321* (0.133)	-0.273** (0.100)	-0.094 (0.193)	-0.164 (0.232)	-0.005 (0.105)	

Table A.3. Full Model Coefficients and Standard Errors (in Parenthesis) Predicting Respondent Behaviors on First Respondent Conversational Turn (continued)

Clarification	Adequate		Problematic substantive		Nonsubstantive		
	Qualified	Uncodable	DK/REF	Don't know	Any clarification request		
Interviewer Controls							
l'er sex (female = 1)	-0.160 (0.110)	0.024 (0.103)	-0.572** (0.196)	-0.257 (0.223)	0.213* (0.104)		
l'er race (white = 1)	-0.064 (0.114)	0.042 (0.107)	-0.007 (0.200)	-0.127 (0.229)	0.264* (0.109)		
l'er tenure (employed 1 + year)	0.128 (0.128)	-0.134 (0.121)	0.122 (0.232)	0.009 (0.266)	0.182 (0.122)		
Constant	0.968* (0.466)	-3.013**** (0.440)	-5.399**** (0.812)	-4.684**** (0.805)	-3.410**** (0.510)		
Variance Components							
SD-Interviewer	0.175	0.139	0.275	0.268	0.137		
SD-Question	0.686	0.601	0.997	0.649	0.715		
SD-Respondent	0.662	0.634	0.787	0.777	0.561		
LR test	1617.27****	589.25****	199.50****	38.95****	619.66****		
Model Fit Statistics							
Log-likelihood	-9971.47	-6234.52	-1392.69	-823.80	-4614.33		
AIC	20002.94	12529.04	2845.38	1707.59	9288.66		
Wald chi-square	256.88****	298.59****	60.47***	70.72****	125.75****		

n = 20,887 in all models. + p < 0.10 ; * p < 0.05 ; ** p < 0.01 ; *** p < 0.001, **** p < 0.0001.

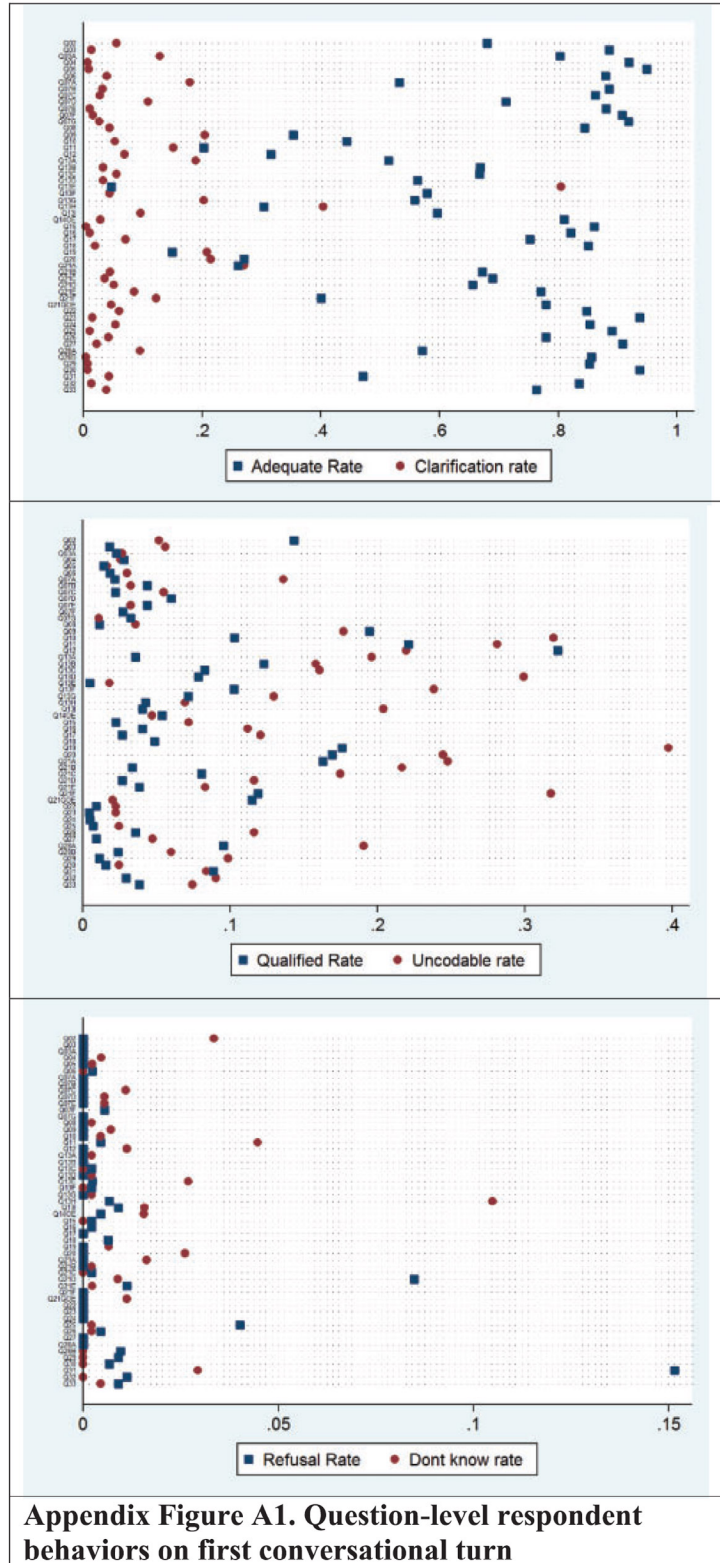


Figure A.1. Question-Level Respondent Behaviors on First Conversational Turn.

Supplementary Materials – Supplementary materials (one .xls file) are archived with this document record.

References

- AAPOR Cell Phone Task Force (2010), “New Considerations for Survey Researchers When Planning and Conducting RDD Telephone Surveys in the U.S. with Respondents Reached via Cell Phone Numbers,” <http://www.aapor.org/Education-Resources/Reports/Cell-Phone-Task-Force-Report.aspx> accessed March 16, 2018.
- Alwin, D. F. (2007), *Margins of Error: A Study of Reliability in Survey Measurement*, Hoboken, NJ: John Wiley & Sons.
- Alwin, D. F., E. M. Baumgartner, and B. A. Beattie (2017), “Number of Response Categories and Reliability in Attitude Measurement,” *Journal of Survey Statistics and Methodology*, DOI: 10.1093/jssam/smx025.
- Alwin, D. F., and B. A. Beattie (2016), “The KISS Principle in Survey Design Question Length and Data Quality,” *Sociological Methodology*, 46, 121–152.
- American Association for Public Opinion Research (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). AAPOR. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf
- Andrews, F. M. (1984), “Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach,” *Public Opinion Quarterly*, 48, 409–442.
- 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.
- Beatty, P., and D. Herrmann (2002), “To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse,” in *Survey Nonresponse*, eds. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 71–85, New York: John Wiley and Sons.
- Belli, R. F., and J. M. Lepkowski (1996), “Behavior of Survey Actors and the Accuracy of Response,” in *Health Survey Research Methods: Conference Proceedings*, DHHS Publication No. PHS 96-1013. Hyattsville, MD: US Department of Health and Human Services. pp. 69–74.
- Beretvas, S. N. (2011), “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.
- Blair, E., and S. Burton (1987), “Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions,” *Journal of Consumer Research*, 14, 280–288.
- Blair, J., and K. P. Srinath (2008), “A Note on Sample Size for Behavior Coding Pretests,” *Field Methods*, 20, 85–95.
- Blumberg, S. J., and J. V. Luke (2010), *Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July–December 2009*, Washington: National Center for Health Statistics.

- Bradburn, N. M., and C. Miles (1979), "Vague Quantifiers," *Public Opinion Quarterly*, 43, 92-101.
- Cannell, C. F., P. V. Miller, and L. Oksenberg (1981), "Research on Interviewing Techniques," *Sociological Methodology*, 12, 389-437.
- Conrad, F. G., and M. F. Schober (2000), "Clarifying Question Meaning in a Household Telephone Survey," *Public Opinion Quarterly*, 64, 1-28.
- Dijkstra, W. (1999), "A New Method for Studying Verbal Interactions in Survey Interviews," *Journal of Official Statistics*, 15, 67-85.
- Dijkstra, W., and Y. Ongena (2006), "Question-Answer Sequences in Survey-Interviews," *Quality & Quantity*, 40, 983-1011.
- Dillman, D. A., J. D. Smyth, and L. M. Christian (2014), *Internet, Phone, Mail, and Mixed Mode Surveys: The Tailored Design Method*, Hoboken, NJ: John Wiley & Sons.
- Dykema, J., J. M. Lepkowski, and S. Blixt (1997), "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study," in *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, pp. 287-310, New York: Wiley.
- Dykema, J., N. C. Schaeffer, D. Garbarski, and M. Hout (in press), "The Role of Question Characteristics in Designing and Evaluating Survey Questions," in *Advances in Questionnaire Design, Development, Evaluation, and Testing*, eds. P. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. Willis, and A. Wilmot, Hoboken, NJ: Wiley.
- Dykema, J., N. C. Schaeffer, D. Garbaski, 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): <http://www.surveypractice.org/article/2817>
- Fail, S., M. F. Schober, and F. G. Conrad (2016), "Hesitation in Socially Desirable Responses in a Mobile Phone Survey," Presentation at the American Association for Public Opinion Research annual meeting, Austin, TX.
- Fowler, F. J. (1992), "How Unclear Terms Affect Survey Data," *Public Opinion Quarterly*, 56, 218-231.
- Fowler, F. J. (2011), "Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions," in *Question Evaluation Methods: Contributing to the Science of Data Quality*, eds. J. Madans, K. Miller, A. Maitland, and G. Willis, pp. 5-21, Hoboken, NJ: Wiley.
- Fowler, F. J., and C. F. Cannell (1996), "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions," in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds. N. Schwarz, and S. Sudman, pp. 15-36, San Francisco: Jossey-Bass.
- Fowler, F. J., and T. W. Mangione (1990), *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Newbury Park, CA: SAGE Publications, Inc.

- Galesic, M., and M. Bosnjak (2009), "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey," *Public Opinion Quarterly*, 73, 349-360.
- Goldshmidt, O. T., and L. Weller (2000), "'Talking Emotions': Gender Differences in a Variety of Conversational Contexts," *Symbolic Interaction*, 23, 117-134.
- Hess, J., E. Singer, and J. Bushery (1999), "Predicting Test-Retest Reliability from Behavior Coding," *International Journal of Public Opinion Research*, 11, 346-360.
- Holbrook, A., Y. I. Cho, and T. Johnson (2006), "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties," *Public Opinion Quarterly*, 70, 565-595.
- Holbrook, A. L., T. P. Johnson, Y. I. Cho, S. Shavitt, N. Chavez, and S. Weiner (2016), "Do Interviewer Errors Help Explain the Impact of Question Characteristics on Respondent Difficulties?," *Survey Practice*, 9(2), <http://www.surveypractice.org/article/2818>
- Holbrook, A. L., J. A. Krosnick, D. Moore, and R. Tourangeau (2007), "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes," *Public Opinion Quarterly*, 71, 325-348.
- Houtkoop-Steenstra, H. (2000), *Interaction and the Standardized Survey Interview: The Living Questionnaire*, Cambridge, England: Cambridge University Press.
- Johnson, T. P., D. O'Rourke, N. Chavez, S. Sudman, R. Warnecke, L. Lacey, and J. Horm (1996), "Cultural Variations in the Interpretation of Health Survey Questions," in *Health Survey Research Methods Conference Proceedings*, ed. R. B. Warnecke, pp. 57-62, DHHS Publication no. (PHS) 96-1013. Hyattsville, MD: National Center for Health Statistics.
- Knauper, B., R. F. Belli, D. H. Hill, and A. R. Herzog (1997), "Question Difficulty and Respondents Cognitive Ability: The Effect on Data Quality," *Journal of Official Statistics*, 13, 181-199.
- Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measurement 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.), eds. P. V. Marsden, and J. D. Wright, pp. 263-313, Bingley, UK: Emerald Group Publishing Limited.
- Lenzner, T. (2012), "Effects of Survey Question Comprehensibility on Response Quality," *Field Methods*, 24, 409-428.
- Lenzner, T. (2014), "Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?," *Sociological Methods & Research*, 43, 677-698.
- Lynn, P., and O. Kaminska (2013), "The Impact of Mobile Phones on Survey Measurement Error," *Public Opinion Quarterly*, 77, 586-605.
- Mathiowetz, N. A. (1998), "Respondent Expressions of Uncertainty: Data Sources for Imputation," *Public Opinion Quarterly*, 62, 47-56.

- 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: John Wiley & Sons, Inc.
- Moore, R. J., and D. W. Maynard (2002), "Achieving Understanding in the Standardized Survey Interview: Repair Sequences," 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. 281-312, New York: John Wiley & Sons, Inc.
- Morton-Williams, J., and W. Sykes (1984), "The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Questions," *Journal of Market Research Society*, 2, 109-127.
- Narayan, S., and J. A. Krosnick (1996), "Education Moderates Some Response Effects in Attitude Measurement," *Public Opinion Quarterly*, 60, 58-88.
- Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents, and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361-396.
- Ongena, Y. P. (2005), "Interviewer and Respondent Interaction in Survey Interviews," Doctoral dissertation, Vrije Universiteit, Amsterdam.
- Ongena, Y. P., and W. Dijkstra (2007), "A Model of Cognitive Processes and Conversational Principles in Survey Interview Interaction," *Applied Cognitive Psychology*, 21, 145-163.
- Ongena, Y. P., and W. Dijkstra (2010), "Preventing Mismatch Answers in Standardized Survey Interviews," *Quality & Quantity*, 44, 641-659.
- Rabe-Hesketh, S., and A. Skrondal (2012), *Multilevel and Longitudinal Modeling Using Stata, Third Edition, Volume II: Categorical Responses, Counts, and Survival*, College Station, TX: Stata Press.
- Raudenbush, S. W., and A. S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.), Thousand Oaks, CA: Sage Publications.
- Saris, W. E., and I. N. Gallhofer (2007), *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, Hoboken, NJ: John Wiley & Sons.
- Schaeffer, N. C., M. Chen, J. Dykema, D. Garbarski, and M. Hout (2015), "Question Characteristics and Item Reliability," paper presented at the Midwest Association for Public Opinion Research annual meeting, Chicago, IL.
- Schaeffer, N. C., and J. Dykema (2011a), "Questions for Surveys: Current Trends and Future Directions," *Public Opinion Quarterly*, 75, 909-961.
- Schaeffer, N. C., and J. Dykema (2011b), "Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions," in *Question Evaluation Methods: Contributing to the Science of Data Quality*, eds. J. Madans, K. Miller, A. Maitland, and G. Willis, pp. 23-39, Hoboken, NJ: Wiley.
- Schaeffer, N. C., and D. W. Maynard (2002), "Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews," 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. 261–280, New York: John Wiley and Sons, Inc.
- Schwarz, N., F. Strack, H. J. Hippler, and G. Bishop (1991), “The Impact of Administration Mode on Response Effects in Survey Measurement,” *Applied Cognitive Psychology*, 5, 193–212.
- Smyth, J. D., and K. Olson (2016), “How do Mismatches Affect Interviewer/Respondent Interactions in the Question/Answer Process?,” paper presented at the annual meeting of the Midwest Association for Public Opinion Research, Chicago, IL, November 18–19, 2016.
- Suchman, L., and B. Jordan (1990), “Interactional Troubles in Face-to-Face Survey Interviews,” *Journal of the American Statistical Association*, 85, 232–253.
- Tourangeau, R., F. G. Conrad, and M. P. Couper (2013), *The Science of Web Surveys*, New York: Oxford University Press.
- Tourangeau, R., L. J. Rips, and K. Rasinski (2000), *The Psychology of Survey Response*, Cambridge, UK: Cambridge University Press.
- 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.
- van der Zouwen, J. (2000), “An Assessment of the Difficulty of Questions Used in the ISSP Questionnaires, the Clarity of Their Wording, and the Comparability of the Responses,” *ZInformation*, 46, 96–114.
- van der Zouwen, J., and W. Dijkstra (2002), “Testing Questionnaires Using Interaction Coding,” 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. 427–448, New York: Wiley.
- Velez, P., and S. D. Ashworth (2007), “The Impact of Item Reliability on the Endorsement of the Midpoint Response in Surveys,” *Survey Research Methods*, 1, 69–74.
- 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.