

Time after Time: Effects of Interview Setting, Interviewers, Respondents, and Items on Response Times

Daniel Goldstein¹ and Elyzabeth Gaumer¹

¹Center for Research on Housing Opportunity, Mobility and Equity (HOME),
Department of Housing Preservation and Development, City of New York,
100 Gold Street, New York, NY 10038*

Abstract

Questionnaire length is an important factor in survey cost and response rates. Reducing respondent burden through improved questions can potentially result in higher data quality and even prevent breakoffs. Advances in survey technology allow the measurement of overall interview length as well as precise capture of response times to individual questions. An emerging literature has examined the effects on response times of various item features as well as respondent and interviewer characteristics in a multilevel analytic framework in web, in-person, and telephone surveys. This paper analyzes in-person interviews from the New York City Housing and Neighborhood Study, a randomized natural experiment evaluating the impact of affordable housing on overall wellbeing. Householder interviews (affordable housing applicants with no co-resident children) took on average 60 minutes; caregiver interviews (applicants with co-resident children) took on average 90 minutes and included all householder questions plus additional caregiving and child-related questions. Interviews were conducted variously in respondents' homes, at project offices, and in public locations like coffee shops, libraries, or parks. This paper builds on previous work by examining the variability in response time due to interviewers, respondents, and items. We explore which characteristics of each are associated with response time as well as investigating interview setting through a sociocultural as well as a cognitive perspective. Results provide insight into which interview settings may be more encouraging to respondents as well as which question types and answer features pose challenges, and which may be easier to answer. This analysis supports improvements in study design that may lower respondent burden and fatigue, produce higher data quality, and reduce break-offs.

Key Words: interviewer effects, response times, questionnaire design, multilevel model

1. Introduction

There are innumerable decision points involved in writing and ordering survey items for a questionnaire. Survey methodologists tend to focus on measurement properties and analytic needs, while survey operation centers are more grounded in the practical cost-benefit realities of data collection as measured in dollars-per-minute or -per-data-point. Analysis of response times has been used to support these aims but can also be used in thinking about survey design more holistically. A complete perspective involves centering the respondent and their experience throughout an interview—"a conversation with a purpose" (Bingham & Moore, 1934/1959). In this broad view, response times can be a lens on the contextual effect of the whole interview—how the respondent cumulatively experiences the ebb and flow of effort and time that they are asked to devote to successive questions during the interaction and how survey teams can leverage this to optimize among their various goals.

In this paper, we examine how survey response times vary by interview location, respondent and item characteristics, and substantiveness of the answer. We further assess how much of the overall variance in response times can be explained by the interviewer asking the question, the respondent answering the question, and the question itself. We replicate the analytic approach used by others to study response times and apply it to about 390,000 response times generated from the New York City Housing and Neighborhood Study (NYCHANS), a 60-90-minute interview conducted in-person with low-income adults ($n=1,524$ respondents). We address two research questions. First, what relative contributions do items, respondents, and both interviewers make to the variability in these response times? Two, what features or characteristics of interview location, questions, response options, and respondents are associated with faster or slower response times? We discuss our results within the context of the existing evidence base and provide suggestions for next steps in response time research.

2. Background

Response times are simply the amount of time that it takes a respondent to answer an interview question. Analyses have generally focused on one of two goals: data collection efficiency and data quality improvement. The more practical strain of research uses response time measurement to estimate survey length and plan data collection (Olson & Parkhurst, 2013; Schneider et al., 2023; Wenz, 2015; and see Loewen et al., 2022). At the most complex, Sturgis et al. (2021) suggest modeling question and answer characteristics to predict response times for new survey items and summing across to estimate total interview time.¹ More often, response times serve as a convenient, though imperfect, proxy for respondent focus or lack thereof and corresponding low-quality responses—an indirect measure of data quality that is particularly useful because it is observed for every item in every interview (Garbarski et al., 2020; Olson & Smyth, 2015; Sturgis et al., 2021). Specifically, response times have been widely used to identify problematic items and detect speeding and satisficing (Andreadis, 2021; Heerwegh, 2002; Krosnick, 1991; Tourangeau et al., 2004; Zhang & Conrad, 2014). While response times that are too fast can indicate lack of respondent attention or effort, slower response times are associated with incorrect knowledge and may indicate longer cognitive processing by a respondent (Callegaro et al., 2009; Kaminska et al., 2010; Malhotra, 2008; Bassili & Krosnick, 2000; Heerwegh, 2003; Turner et al., 2015). This line of research has been extended to explore the association between interviewer and respondent behaviors that indicate poor data quality and elements of response times like longer response latencies and question reading (Draisma & Dijkstra, 2004; Olson et al., 2019). More recent work has aimed at marrying response times to data quality more directly by analyzing them alongside item nonresponse in the form of refusals or “don’t know” responses (Dahlhamer et al., 2020).

2.1 Response Times and the Cognitive Response Process

Response time are almost universally analyzed through a cognitive lens, starting with their original deployment as a dependent variable to illuminate the cognitive process through attitude formation and association strength (Bassili & Scott, 1996). In recent years, survey researchers have reversed the equation, exploring the cognitive determinants of response times (Yan & Tourangeau, 2008). Each of the four elements of Tourangeau et al.’s (2000) seminal model of the response process—understanding the question, retrieving relevant information, forming judgments, and providing a response—may contribute to shorter or longer response times, and each may be influenced by respondent or item characteristics in any interview mode (Yan & Tourangeau, 2008). Among other factors, *comprehension* is influenced by both complexity of the question and cognitive ability of the respondent. *Recall* relates to both the salience of the topic and working memory of the respondent. *Judgment* depends on the framing of the question and context established by previous items as well as the knowledge of the respondent. *Reporting* decisions are affected by the number and type of answer choices to select from as well as a given respondent’s tendency toward overthinking or

¹ The level of sophistication and reliability in timing approximation varies, with some online platforms recently incorporating proprietary machine learning algorithms into products to “score” surveys and estimate completion time.

precision. Interviewer administration can intervene on each stage of the process, potentially speeding or slowing the cognitive processes shaping response times beyond a simple respondent-item framework.

2.1.1 Respondent characteristics

Respondent characteristics associated with cognitive capacity have been found to predict response time across various modes. The most robust findings is that response times are slower for older respondents (Couper & Kreuter, 2013; Dahlhamer et al., 2020; Garbarski et al., 2020; Loosveldt & Beullens, 2013; Olson & Smyth, 2015; Sturgis et al., 2021; Wenz, 2015; Yan & Tourangeau, 2008; Zhang & Conrad, 2014). Almost all studies have found that response times are faster for respondents with higher levels of education (Couper & Kreuter, 2013; Garbarski et al., 2020; Wenz, 2015; Yan & Tourangeau, 2008) (but see Dahlhamer et al. (2020) for the surprising opposite result). Employed respondents have been shown to answer faster, as have those with higher incomes, possibly because these characteristics are associated with less mental scarcity and more available cognitive capacity or because these respondents may be more likely to have pressing demands on their time outside of the interview (Dahlhamer et al., 2020; Mullainathan & Shafir, 2013; Olson & Smyth, 2015). Couper and Kreuter (2013) found that Spanish-language interviews in the United States had slower individual response times than English—a finding that parallels differences in overall interview length duration—while elsewhere nativity had no association with response times (Dahlhamer et al., 2020). The slower Spanish-language responses may be explained by the extra cognitive load of multilingual functioning in society in general and the interview setting in particular, or they may relate to often inferior translation quality relative to the original instrument and the resulting additional cognitive burden of understanding the questions and answer choices.

2.1.2 Item characteristics

Item characteristics, whether inhering in the question stem or response option(s), undoubtedly influence how long it takes to answer the question. Features of the question stem primarily influence early stages in the cognitive response process—comprehension and retrieval. Response options function through the later stages of judgment and reporting.

Question features

According well with the cognitive response process model, certain types of questions have been found to be associated with faster responses times than others—demographic questions are generally fastest with items assessing attitudes or subjective opinions slower, in particular where attitudes are not strongly held or have not yet been formed at all, though some studies have found that recall of some behavior questions is slower (Dahlhamer et al., 2020; Olson & Smyth, 2015; Sturgis et al., 2021; Tourangeau et al., 2000; Yan & Tourangeau, 2008). Studies have consistently confirmed the intuitive finding that longer questions have slower response times—more words take more time to hear or read (Couper & Kreuter, 2013; Dahlhamer et al., 2020; Garbarski et al., 2020; Olson & Smyth, 2015; Sturgis et al., 2021; Yan & Tourangeau, 2008). Above and beyond length, more complex questions that demand more cognitive processing—whether measured at a higher reading level or with more clauses per question or words per clause—are associated with slower response times (Garbarski et al., 2020; Olson & Smyth, 2015; Yan & Tourangeau, 2008). Perhaps surprising from a cognitive process viewpoint, Dahlhamer et al. (2020) found that optional text was associated with faster response times, despite additional processing burden on interviewers and additional reading time for some respondents. Items with interviewer instructions were found by Garbarski et al. (2020) and Sturgis et al. (2021) to be slower, presumably because of the interviewer's additional reading and processing requirement, but Couper and Kreuter (2013) found that they were faster, and Olson and Smyth (2015) found no association. Most studies have found that items placed later in the instrument were associated with faster response times, possibly through a combination of respondent learning, interviewers increasing the pace, and respondent expending less effort and instead satisficing (Garbarski et al., 2020; Yan & Tourangeau, 2008), though Couper and Kreuter (2013) found a slowdown in response times as the interview progressed, possibly because of fatigue. A question feature that has not been examined in the literature is looping or iteration—where the same item is repeated about a different subject person, such as for household demographics. We

expect respondents to answer later iterations of the same question faster as they learn what sort of answer is required and begin to anticipate the next question.

Answer features

Beyond the question stem, the cognitive response model has also been applied to response options to examine response times. Generally, fully open-ended answers are associated with slower response times, and yes/no answer options are among the fastest (Couper & Kreuter, 2013; Dahlhamer et al., 2020; Garbarski et al., 2020; Olson & Smyth, 2015). An increase in the number of response options is associated with slower response times, likely because of the extra effort in choosing among additional possibilities (Olson & Smyth, 2015; Yan & Tourangeau, 2008). In a similar effect, questions whose answer choices were shown to respondents on a card were associated with slower response times (Couper & Kreuter, 2013). Slower responses when presented a flashcard or response card, however, may not reflect the card itself, but rather the additional cognitive burden of more complicated or numerous answer choices for which these cards are usually created.

We would expect respondents to answer questions that are linked in some way to earlier items more quickly because they have already activated the necessary memory structure (Tourangeau et al., 2000), and response times for topically linked questions that share answer choices—a battery—should be faster still because respondents do not need to redo the work of understanding answer choices for each question. In some dyads, an interviewer may save time by not even reading the identical question stem to the respondent for later items in the battery. Indeed, Sturgis et al. (2021) found that, relative to non-battery questions, later items in a battery of questions were associated with faster response times, while the first time was associated with slower response times. But other research found no effect of being in a battery on response times (Garbarski et al., 2020; Olson & Smyth, 2015).

2.2 Response Times and Sociocultural Factors

Bingham and Moore's (1934/1959) early description of the research interview as a "conversation with a purpose" holds as true for survey as for qualitative interviews, though the emphasis is squarely on the purpose rather than the conversation. Much has changed in survey research in the intervening decades, but the fundamental truth that an interview is a social interaction—and therefore embedded in sociocultural dynamics—has not. Miller (2004) has demonstrated that the cognitive model is not sufficient; rather, social and cultural factors must be considered to explain the survey response process and corresponding response times more fully. For example, varying conversational norms across languages and cultures may dictate longer pauses between turn-taking or, conversely, encourage interrupting. Certain cultural orientations may encourage a view of interviewers as authority figures deserving of deference or distrust resulting in longer pauses; respondent social location vis-à-vis the interviewer or study sponsor may have a similar impact. This perspective may help to explain influences on response times (and measurement error more broadly) of characteristics unrelated to the cognitive process or heavily influenced by non-cognitive factors. For respondents this includes race/ethnicity, nativity, language spoken, and gender, and for items, perceived sensitivity or invasiveness and location in the instrument as rapport wanes or strengthens.

2.2.1 Interviewers

In line with a broader view of the response process, Tourangeau (2018) conceptualized interviewer contribution to response as more in line with a conversational than cognitive model. There is broad recognition that interviewers are a source of survey error through measurement (separate from representation through recruitment and non-response) and that their contributions vary both by item, respondent, and interviewer characteristic (Schaeffer et al., 2010; West & Blom, 2016). Pirralha, Haag, and Maurice (2023) found that interviewers account for about 30% of the variation in overall length of a telephone interview, generally according with in-person interview estimates. At the level of survey module, interviewers across countries in the European Social Survey determine interview speed, more so than respondents (Loosveldt & Beullens, 2013). At the level of individual response times, the focus of this paper, the literature generally shows a small but statistically significant

contribution of interviewers to response time variability (Dahlhamer et al., 2020; Garbarski et al., 2020; Olson & Smyth, 2015), with more experienced interviewers associated with faster response times (Couper & Kreuter, 2013; Sturgis et al., 2021). The cognitive response model would posit that increased experience accelerates interviewer processing of instructions and decision-making around optional text (Bergmann & Bristle, 2020). But competing explanations—rooted in the social interaction between interview and respondent—would point to the tendency of many interviewers not to read verbatim, especially where these protocol deviations vary by respondent characteristics (Bell et al., 2016; Yu et al., 2019).

2.2.2 Interview setting

The physical setting and broader location of an interview contributes profoundly to the cultural dynamics of the social interaction, potentially influencing item response times. Herzog (2012) has argued that qualitative interview locations should not be considered a matter of convenience or happenstance but rather conceptualized as a fundamental feature of the data collected. Selection of a location reflects not just the power dynamics in the purposeful conversation but contributes to the construction of the reality of the interview and should be considered part and parcel of the findings. Similarly, Leverentz (2023) contends that the location of a qualitative interview shapes its dynamics. But these insights apply equally to structured face-to-face survey interviews. Whether an interview takes place in the respondent's home or workplace, an officially designated survey project space, or a neutral, third location may influence respondent and interviewer levels of comfort, engagement, and motivation. The respective social locations that the interviewer and respondent occupy outside of the interview remain regardless of the location and the fact of the interview, but the location may also color how each inhabit the social roles of “interviewer” and “respondent” (Schaeffer, 2004).

Location is of course most relevant for face-to-face interviews—the focus of this study—though respondent attention and focus may vary depending on where they answer an interviewer's phone call, complete a web survey, or respond to questions unassisted in a computer-assisted self-interview (CASI) portion of face-to-face interview. Significant contributions of mode effects to answer variability have long been recognized (Dillman & Christian, 2005; Schouten et al., 2021), and accordingly, response times differ across modes, with audio computer-assisted self-interview (ACASI) responses provided faster than CASI, and in-person responses faster than telephone (Couper & Kreuter, 2013; Dahlhamer et al., 2020).

Couper (1996) exploited the shift from paper-and-pencil interviewing (PAPI) to computer-assisted personal interviewing (CAPI) in the Current Population Survey (CPS) to explore the impact of interview setting and location, specifically inside the respondent's home compared to elsewhere (mainly on the doorstep), finding that total interview times were longer in the home. To our knowledge, however, no study has explored the impact on interview location on individual item-level response times.² We compare among three types of locations: in-home interviews, those conducted at project offices, and those conducted at other neutral locations such as cafes, parks, and the like. Though changing locations may tweak cognitive processing speed or travel to various locations may increase fatigue for interviewers or respondents, location likely influences response times more broadly because the setting determines the ground rules and baseline (dis)comfort that may speed up or slow down responses. Some respondents may find the office intimidating or uncomfortable, while others may be uncomfortable with interviewers in their home. Additionally, there may be more distractions at home or neutral location relative to project offices.

3. Data and Methods

3.1 Data

The data for this paper come from the New York City Housing and Neighborhood Study (NYCHANS), an experimental study evaluating the impact of affordable housing on the well-being

² Location should not be confused with area, which should often be considered when analyzing interviewer effects because of its conflation with interviewer in interviewer assignment procedures in many surveys (see Sturgis et al. (2021) and Campanelli and O'Muircheartaigh (1999)).

of low-income New Yorkers (Goldstein & Gaumer, 2022). Respondents were members of households that applied to affordable housing and qualified for one or more housing unit. Data collection was undertaken by an in-house field team at the New York City Department of Housing Preservation and Development in-person primarily via CAPI from 2014 through 2018 (see Waickman et al. (2019)). Interviews were administered by trained field interviewers with a careful choreography that required two interviewers for each adult respondent in roles titled first and second chair, together designed to ensure respondent comfort as well as data quality. Responsible for reading questions and capturing responses, first chair interviewers monitored interview pace. Second chair interviewers functioned as a support system for respondents, staying attuned to their physical, mental, and emotional needs and deploying body language and other soft skills and to help them remain at ease.

NYCHANS collected data on 1,654 respondents through interviews with adults across two overlapping instruments, the householder interview for those who applied for housing without a co-resident child and the caregiver interview for those who applied with one or more co-resident child (AAPOR (2016) RR1 71.5%). Analysis of response times here are based on 1,524 cases and conducted on unweighted data; however, baseline characteristics for treatment and controls were balanced, and response rates were comparable. The householder instrument for adults who applied without children comprised about 300 questions and averaged 60 minutes. Topics covered include household roster and residential history; income and employment; housing costs and condition; neighborhood disorder, quality, safety, collective efficacy, and amenities; physical and mental health and health behaviors; and financial wellbeing. All items were replicated exactly in the instrument for caregivers with children, with some modules randomly assigned for removal to make room for about 100 additional questions about childcare, child health and wellbeing, and caregiving related topics. The more comprehensive caregiver interview averaged 90 minutes. CASI modules for potentially sensitive questions about child behavior were included in the caregiver interview but are not included in this analysis. An abbreviated CASI version of the instrument was deployed as a last-resort interview; these cases were not included in this analysis.

Because of the cell size requirements for analyzing data with the complex, multiply nested nature described below, interviews were excluded if the first chair interviewer conducted fewer than five interviews in that role (26 cases) or the second chair conducted fewer than five interviews in that role (45 cases). Interviews longer than three hours were excluded, which includes those that were interrupted and resumed later and those with a timing or other processing glitch (8 cases: min 970 minutes, max 595 days). Interviews were also excluded where complete information about respondents was not available because of non-response (51 cases).

3.1.1 Response times (dependent variable)

The dependent variable for this analysis is a log-transformation of the total number of seconds spent on each item, referred to as response times. Here, reflecting our conceptualization of the survey interview as a fundamentally social interaction, response times are expansive. They include the time taken by interviewer reading and computer navigation that partially overlaps with the time it takes a respondent to hear and understand the question and conceptualize and verbalize their answer as well as the time for the interviewer to record it. All CAPI portions of the survey relied on SurveyToGo, a commercially available computer-assisted interviewing (CAI) tool developed by Dooblo. The CAPI software produced paradata about the interview, including timestamps (to the second) for the final time an interviewer entered the answer to a question. These timestamps were extracted alongside survey responses. The response time attributed to each item was calculated by subtracting the timestamp of the previous item. For example, if Question 5 was answered at 1:00:00 PM and Question 6 was answered at 1:00:10 PM, the response time for Question 6 would be 10 seconds.

Data procedures relating to exclusions, handling of outliers, and transformation of response times were performed before the exclusion of entire interviews for analytic purposes described above. The calculation of a response time for a given item depends on the validity of both its own timestamp and that of the previous item. We therefore exclude items where the timestamp for an item or the

preceding item is not well-defined (22 items). This includes items directly after items involving interactive interview components, such as capturing residential history using an interview card. At times the respondent or interviewer backtracked to change an answer to a previous question after answering a later one, resulting in items where the timestamp for the preceding item was chronologically later than the timestamp for the item that was later in the instrument. These response times could not be computed and were excluded (18,509 responses). Questions to which respondents did not know or refused to provide the answer *were included* because interviewer and respondent behaviors in response to these common survey interactions are substantively meaningful. But items where the interviewer chose to skip asking the question because of time constraints or extreme discomfort of the respondent with earlier questions in the module were excluded (2,210 response times). Response times that were implausibly short or long were also excluded. Although some questions were extremely short and could be answered very quickly, response times of 0 seconds were excluded (774 response times). Response times over five minutes, which we considered indicative of an outside distraction or some process unrelated to survey response, were excluded (105 response times). Response time outliers were defined at the question level as the 1st and 99th percentiles. Observations below or above these cutoffs were replaced with the percentile values, following standard practice (Olson & Smyth, 2015; Ratcliff, 1993; Yan & Olson, 2013; Yan & Tourangeau, 2008).

Our data were highly skewed, a distribution typical of response times (see Figure 1). Following standard practice, we transformed the trimmed individual response times using the natural log.

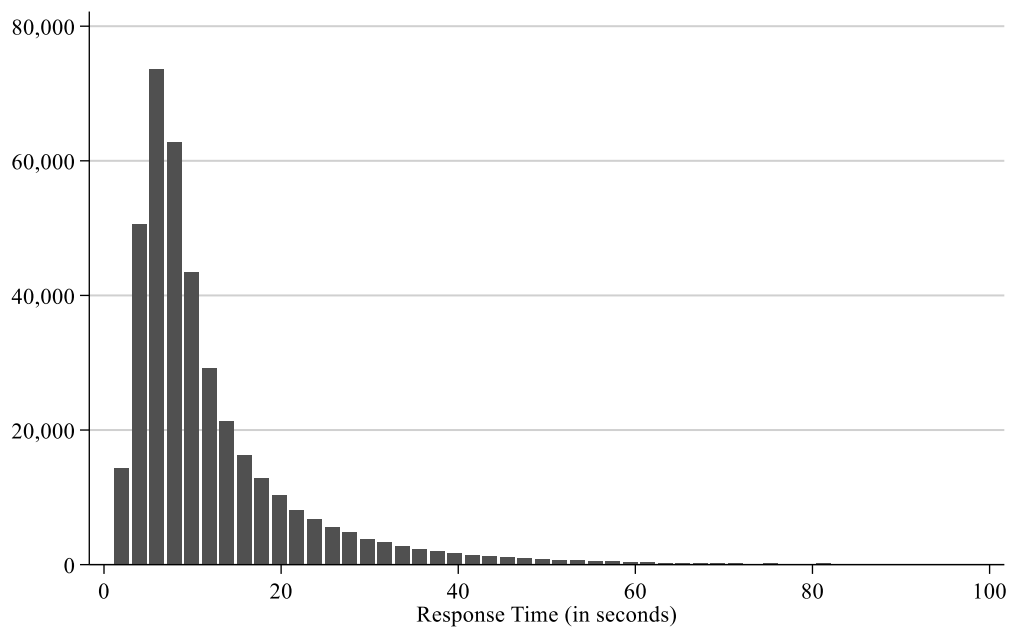


Figure 2: Histogram of trimmed response times in seconds (879 response times (0.2% of total) longer than 100 seconds not shown).

3.1.2 Measures (*independent variables and controls*)

A variety of independent variables are included in this analysis, including measures capturing information about the setting of the interview, the respondent, each item's question and answer characteristics, and the response itself. Table 1 provides counts and percents or means and standard deviations, where appropriate, about responses, interviewers, interview locations, and respondents. Where the interview took place was coded into one of three locations: in-home, in-office, or another location. NYCHANS interviews with minor-age children were only conducted in the project's offices where additional protections and oversight were possible. For this reason, caregiver interviews were primarily, though not exclusively, conducted in-office. Whenever possible,

householder interviews were conducted in the respondent's home; however, all respondents were able to elect to do the interview in the project's offices. Respondents who preferred a location other than their home could work with the recruitment team to identify another mutually agreed upon place, such as a park, building lobby, or coffee shop. In all cases, interviews were only conducted where privacy could be assured, including distance from other people. In the present analysis, two dummy variables are included for in-office interviews and other-location interviews.

Table 1: Descriptive Statistics for Response, Interview, Interviewer, and Respondent Characteristics

	n	%
Responses	387,434	100%
Average interview length (minutes)	72.7 (19.7)	
Average # of questions answered	279.1 (44.1)	
Average time per question (seconds)	12.06 (12.80)	
Average natural log time per question	2.17 (0.75)	
Refused	899	0.2%
Don't know	2,011	0.5%
Interviewers	50	100%
Average # of interviews, 1st Chair	35.88 (40.04)	
Average # of interviews, 2nd Chair	38.88 (51.19)	
Location	1,524	100%
In-home	673	44%
In-office	705	46%
Other location	146	10%
Respondents	1,524	100%
Less than HS	82	5%
Age 50+	378	25%
Female	1,098	72%
Employed	1,144	75%
Speaks a language other than English	787	52%
Interview in a language other than English	96	6%

While most analysis treats missing data for a given item as a singular value, either excluding the case or imputing data using one of a variety of strategies, here we coded non-responses to individual items separately for “don’t know” and “refused.” These are compared to non-missing, substantive responses as the reference group.

Respondent characteristics were captured during the NYCHANS interview and included a range of socio-demographic characteristics examined here. Respondent's education level was coded dichotomously as having less than a high school degree (5%), either in the form of a diploma or GED. Age was coded dichotomously as 50 or older (25%) based on the respondent's age at the time of the interview. Gender was coded dichotomously for those who identified as female (72%) compared to any other gender. Speaks a language other than English (52%) was coded dichotomously for any reported language other than English, even if the respondent also spoke English. Interviews that were completed in a language other than English (6%) were coded based on interview paradata. Where possible, NYCHANS completed interviews in the respondent's preferred language and were available in English, Spanish, Russian, Polish, Mandarin, Cantonese, French, Korean, Arabic, Fulani, and Krio.

Question and answer characteristics were coded by the authors based on the final programmed questionnaire. Table 2 shows the distribution of each characteristic over each of the 346 items. The number of words in each question—including introductory and other required text—was counted and then centered by subtracting from the average number of words across all items in the questionnaire. The same process was used to calculate the centered number of characters per word in the question.

Table 2: Descriptive Statistics for Question and Answer Characteristics

	n	%
Questions	346	100%
Average number of words per question	15.27 (8.44)	
Average number of characters per word	4.58 (0.61)	
Chapter introduction	28	8%
Optional question text ("if necessary")	47	14%
Interviewer instructions provided	20	6%
Follow-up question about person named by R	53	15%
Looped question	78	23%
Battery items	125	36%
Type		
Demographic	30	
Behavior	145	42%
Factual	54	16%
Opinion	117	34%
Answers	346	100%
Binary (yes/no or true/false)	99	29%
Closed ordinal	134	39%
Closed nominal	47	14%
Open-ended numerical	48	14%
Open-ended text	18	5%
Select all that apply	18	5%
Average number of response options	6.04 (2.47)	
Scale with all choices labeled	95	27%
Scale with only end points labeled	38	11%
List of people R named are response options	16	5%
Response card aid	159	46%

Additional questionnaire features that were not directly part of the question or answer were also coded. When there was an introductory or transition script that preceded a set of questions, the first question in that section was flagged as having a chapter introduction. Optional interviewer text, such as a definition of a term or alternate phrasing of a concept, was flagged as a characteristic of the item itself, as were any interviewer instructions that were programmed in the instrument but not intended to be read to the respondent, such as an indicator to round an answer to the nearest dollar or to enter height as feet in one field and inches in the next.

Some items in the NYCHANS questionnaire were looped (repeated) for multiple people (e.g., demographics of each household member or frequency of contact with individuals named in the social network module, "Using Response Card A, how often do you talk to [contact]?"). For example, the household roster began by asking the age of person 1 ("How old is [Person 1]?"), followed by relationship ("What is [Person 1]'s relationship to you?"), and so on. After the last

question in the loop, the interviewer would ask about the age of person 2, (“How old is [Person 2]?”) followed by each successive question in that loop. The number of iterations depended upon the size of the household or network. Here, questions in a loop were flagged and categorized according to its iteration in the interview: the first iteration, the second iteration, or the third or later iteration. Some items in the questionnaire were part of a battery, with shared question structure and answer choices. For example, NYCHANS replicated several items on neighborhood disorder where the interviewer asked, “How much of a problem is _____?” with answer choices of “a big problem,” “somewhat of a problem,” and “not a problem.” The respondent was asked the set of questions one after the other. In this analysis, items that were part of a battery were flagged and coded according to whether it was the first question in the battery or the second or higher in that same set.

Questions were sorted according to the order in which they were asked and numbered 1 through k , then centered by subtracting the number order from the overall average count of items across all interviews. To make the coefficient easier to interpret, this was then transformed and presented in hundreds (i.e., centered order number / 100).

Answer characteristics were coded in several ways. This included the number of response choices as well as whether it was a select-all-that-apply as opposed to a forced single response option. Types of answer choices were coded into one of several categories: closed ordinal (such as Likert-type scales), binary (e.g., Yes/No or True/False), closed nominal (“Using Response Card A, what is your race?” with answer options of “A. White, B. Black or African American, C. Asian, D. American Indian or Alaska Native, E. Native Hawaiian or Other Pacific Islander, F. Other”), or open-ended numerical (“How much is your monthly rent?”).

Interview aids were employed in various ways during the NYCHANS interview. Some items referenced a list of individuals from a card, which had been filled out based on answers to earlier questions. An example would be the list of people who live in the household, also termed the household roster card. In other cases, questions asked for respondents to use a filled-out card as list of possible select-all-that-apply answers. For example, to efficiently collect information about household members, respondents were asked, “Who is female?” and then identified female household members on the roster card. Nearly half of the NYCHANS questions referenced a show or flash card that provided a written set of the answer options that could be referenced. Items that employed one or more of these aids were flagged as such.

NYCHANS utilized several different types of scaled answers, including Likert-type scales and numeric ratings. These were coded according to the use of labels: those that labeled only end points and those that labeled each value. For example, some questions asked for respondents’ level of agreement with statements about their neighborhood. Respondents were presented with answer cards with five choices fully labeled: “A. Strongly agree, B. Agree, C. Neither agree nor disagree, D. Disagree, E. Strongly disagree.” These types of answers were categorized as All choices labeled. Some questions asked for level of agreement on a scale from 1-10 and presented an answer card with each value from 1-10 listed from left to right and labeled “< Disagree” above the 1 and “Agree >” above the 10. These were categorized as Only end points labeled.

NYCHANS collected a variety of information across many domains. Some questions asked for basic demographic information. Others asked respondents to recall their behaviors (e.g., “How many times do you or someone else who lives with you walk the dog(s) on a typical day?”), to provide factual information (e.g., “Do you rent or own your home?”), or to provide their subjective opinion about their housing, neighborhood, or something else (e.g., “Using Response Card B, how would you rate the overall physical condition of your current housing?”).

Controls were included in fixed effects models. Household size and presence of a child in the household were controlled for because they determined the number of loops and contributed substantially to the overall length of the interview. Caregiver interview was included to control the effects of different overall instruments and interview procedures.

3.2 Analytic Strategy

The structure of our data is complex, represented graphically in Figure 2. In the present analysis, we consider five classifications: (1) individual responses, (2) items, (3) respondents, (4) first chair interviewers, and (5) second chair interviewers. Because we are interested in both item and respondent contributions to response time and because each item is answered by many respondents and each respondent answers many items, responses are cross-classified by survey items and respondents. In turn, we are also interested in the separate contributions of the interviewers acting as first and second chairs, so respondents are further cross-classified by the two interviewers. The total analytic sample includes 387,434 individual item responses (level 1), including times that form the unit of analysis, to 346 unique items (level 2, classification 2) asked to 1,524 respondents (level 2, classification 3) by one of 46 first chair interviewers (level 3, classification 4) accompanied by one of 42 second chair interviewers (level 3, classification 5).

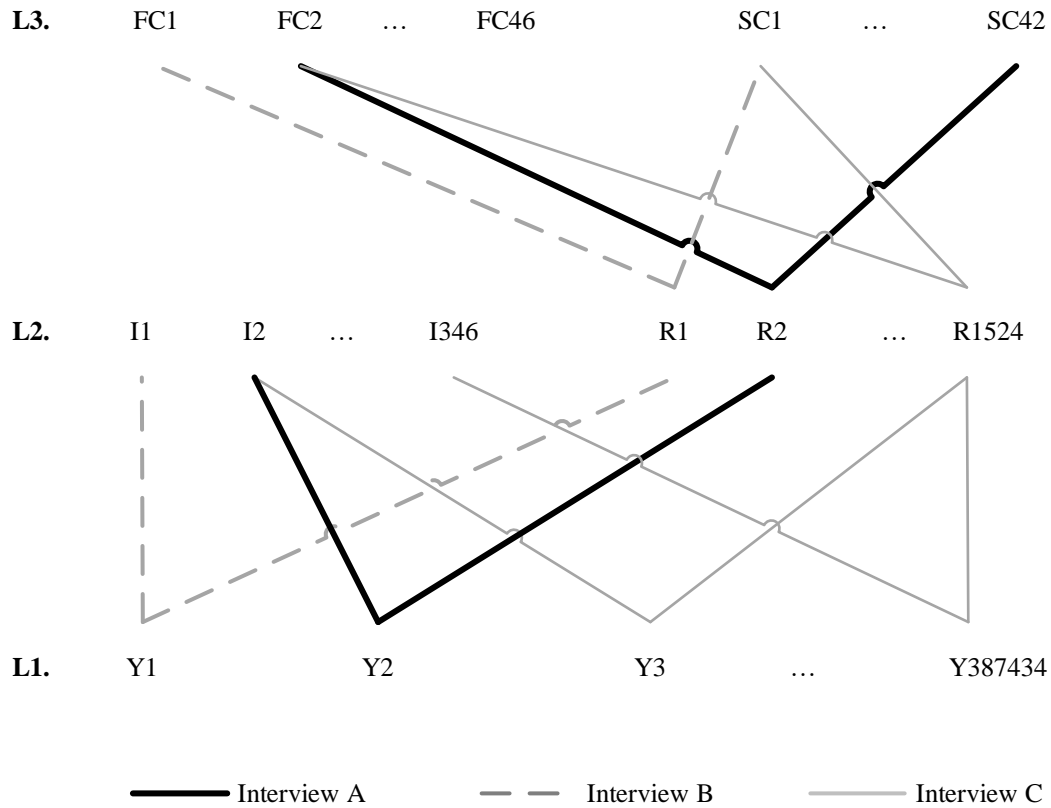


Figure 2: Illustrative example of the data structure for selected responses (Y) cross-classified by items (I) and respondents (R), which are cross-classified by first chair (FC) and second chair (SC) interviewers.

We diagrammed parts of three possible interviews as exemplars in Figure 2. In Interview A, Respondent 2 was asked Item 2 in the questionnaire and provided Response 2. Respondent 2 was interviewed by First Chair 2 and Second Chair 42. In our data structure, Response 2 is cross-classified by Item 2 and Respondent 2. Respondent 2, in turn, is cross-classified by First Chair 2 and Second Chair 42. In Interview C, Response 3 is also cross-classified by Item 2 but with Respondent 1524. In Interview B in the figure, Response 1 is cross-classified by Item 1 and Respondent 1. Each of our 387,434 responses are categorized in this way into one of the 346 items and separately into one of the 1,524 respondents. Each of these respondents is in turn categorized into one of the 46 first chair interviewers and separately into one of the 42 second chair interviewers.

Adapting terminology from Beretvas (2011), our first, base model specification is a three-level cross-classified random effects model with two hierarchies and cross-classification at both level 2 and level 3. All models were estimated using maximum likelihood estimation³ via the *mixed* command in Stata MP 17.0. The base model estimated random intercepts for items, respondents, first chair interviewers, and second chair interviewers (Rabe-Hesketh & Skrondal, 2022). Succeeding models did not include random intercepts for first and second chair interviewers.

Following notation by Browne, Goldstein, and Rasbash (2001), the base model predicts y_i , the natural logarithm of response time for response i ($i = 1, \dots, N$), as a function of an overall mean β_0 , a random effect due to the item $u_{\text{ITEM}(i)}^{(2)}$ (the contribution of item i averaged over all respondents), a random effect due to the respondent $u_{\text{RESP}(i)}^{(3)}$ (the contribution of respondent i averaged over all items), a random effect due to the first chair interviewer $u_{\text{FIRST}(i)}^{(4)}$ (the contribution of first chair interviewer i averaged over all second chair interviewers), a random effect due to the second chair interviewer $u_{\text{SECOND}(i)}^{(5)}$ (the contribution of second chair interviewer i averaged over all first chair interviewers), and a residual term e_i for the i th response in the dataset. Random effect terms are assumed to be normally distributed with mean zero and variance $\sigma_{u(2)}^2$, $\sigma_{u(3)}^2$, $\sigma_{u(4)}^2$, and $\sigma_{u(5)}^2$, respectively, and e_i is normally distributed with mean zero and variance σ_e^2 as given in Equation 1:

$$\begin{aligned} y_i &= \beta_0 + u_{\text{ITEM}(i)}^{(2)} + u_{\text{RESP}(i)}^{(3)} + u_{\text{FIRST}(i)}^{(4)} + u_{\text{SECOND}(i)}^{(5)} + e_i \\ u_{\text{ITEM}(i)}^{(2)} &\sim N(0, \sigma_{u(2)}^2) \\ u_{\text{RESP}(i)}^{(3)} &\sim N(0, \sigma_{u(3)}^2) \\ u_{\text{FIRST}(i)}^{(4)} &\sim N(0, \sigma_{u(4)}^2) \\ u_{\text{SECOND}(i)}^{(5)} &\sim N(0, \sigma_{u(5)}^2) \\ e_i &\sim N(0, \sigma_e^2) \end{aligned} \quad (1)$$

The unconditional base model with random terms for all four classifications partitions the total variance in response times among between-cell (for items, $\sigma_{u(2)}^2$; respondents, $\sigma_{u(3)}^2$; first chair interviewers, $\sigma_{u(4)}^2$; second chair interviewers, $\sigma_{u(5)}^2$) and within-cell (residual, e_i) components. Using these variances to calculate intra-class correlation coefficients (ICCs), following Equation 2, allows us to estimate the proportion of variance in log of response times stemming from each source. ICCs are calculated as the ratio of the variance for one classification to the total variance. For example, the intra-item correlation coefficient (the ICC for items)—calculated as the ratio of the between-items variance to all variance—is interpreted as the share of the variation in response times attributable to differences among items, as given in Equation 2:

$$\rho_{\text{ITEM}} = \frac{\sigma_{u(2)}^2}{\sigma_{u(2)}^2 + \sigma_{u(3)}^2 + \sigma_{u(4)}^2 + \sigma_{u(5)}^2 + \sigma_e^2} \quad (2)$$

We modify Equation 2 with the appropriate variance terms in the numerator for the intraclass correlations of respondents, first chair interviewers, and second chair interviewers, respectively. For simpler versions of our model where interviewer variances are not modeled separately (Models 1, 2, 3, and 4), the variance terms for the first and second chair interviewers are removed from the denominator.

To assess the multilevel structure of our data, we compared the unconditional base model with four classifications (Equation 1) to more parsimonious models. We compared model fit in two ways. One, following Rabe-Hesketh and Skrondal (2022), we performed likelihood-ratio tests after fitting models with different random effects terms. Two, we used Akaike information criterion (AIC) and

³ This follows Angrist and Pischke's (2009) rule of thumb, given our 42 second chair interviewers, the smaller of the two top-level clusters.

Bayesian information criterion (BIC). Our model of best fit (Model 1) removed the two random interviewer terms (classifications 4 and 5 at level 3), leaving random effect terms for only items and respondents (classifications 2 and 3 at level 2). We then added item-level and respondent-level predictors as fixed effects,⁴ as well as various controls, to the simplified unconditional model to identify characteristics associated with faster or slower response times:

$$y_i = \beta_0 + x_1\beta_{MISS} + x_2\beta_{LOC} + x_3\beta_{RESP} + x_4\beta_{ITEM} + u_{ITEM(i)}^{(2)} + u_{RESP(i)}^{(3)} + e_i \quad (3)$$

where x_1 is a vector of response-level covariates related to substantive responses with coefficients β_{MISS} , x_2 is a vector of interview-level covariates related to location with coefficients β_{LOC} , x_3 is a vector of respondent-level covariates with coefficients β_{RESP} , and x_4 is a vector of item-level covariates with coefficients β_{ITEM} . All models additionally control for the longer caregiver interview versus the shorter householder interview, household size, and the presence of a child in household.

4. Results

4.1 Random Effects

Figure 3 shows the ICC results from the null model presented in Equation 1, the model that contains no covariates and merely partitions variance among the five classifications: first chair interviewer, second chair interviewer, respondent, item, and the residual variability in response times that remains unexplained by the random effects in the model. The ICCs indicate that 1.9% and 0.4% of the variability in response times are due to each interviewer respectively, 4.9% is due to the specific respondent, 52.7% is due to the specific item, and 40.2% remains unexplained by interviewers, respondents, or questions.

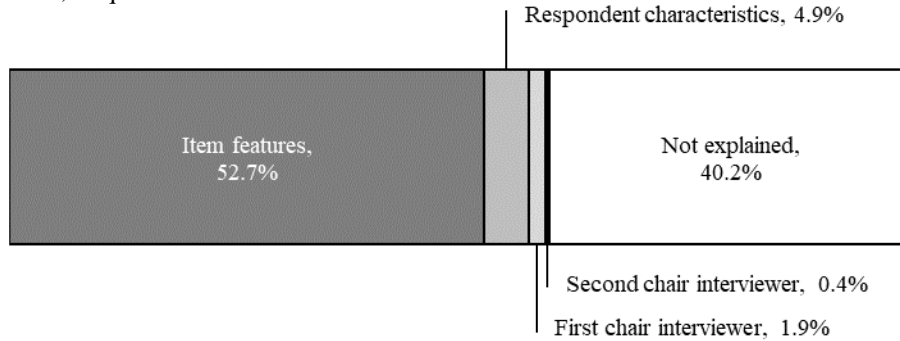


Figure 3: Partition of the variance in response times among random effect terms (Equation 1)

The ICC for items can be interpreted as the average expected correlation in log response times for the same item for two randomly selected respondents and sets of interviewers. In this null model—averaged across all items, respondents, and interviewers, respectively—response times for a given item are more than 10 times more similar than response times for a given respondent, and 28 times more similar than response times for a given interviewer. Put another way, the differences in response times between items are much larger than the differences between respondents, which are larger than the differences between interviewers. This aligns with an intuitive understanding of an instrument with widely varying item type and length. Some question stems only include a few words and require a yes-or-no answer, while some contain dozens of words and request an open-ended response. Variation between these responses times is large, and the questionnaire contained a variety of questions in between. The fastest and slowest respondents, however, were much closer together, and most bunched around the middle.

⁴ Fixed effects terms were added iteratively to the base model (Model 1). Model 2 included terms related to substantive response and location; Model 3 added respondent terms, and Model 4 added item terms.

Because the variance contribution from each interviewer term was very small, we compared the fully specified five-classification base model (Equation 1) with more parsimonious models to determine the most appropriate unconditional null model for examining fixed effects.⁵ The best-fit model, Model 1, contained no random term for interviewers and was limited to a simpler cross-classified model with random terms for only respondents and items.⁶ The ICCs resulting from this model indicated that 51.6% of variability was due to respondents—essentially unchanged from the fully specified null model. 6.7% was due to items, and 41.7% remained unexplained.

4.2 Fixed Effects

Building upon the parsimonious cross-classified model, we introduced a series of predictors as fixed effects to assess their expected associations with response times. The covariates were measured at the response level, the respondent level, and the item level, respectively. Each of these models include controls described above: household size, presence of one or more child in household, and whether the interviewer was a caregiver interview with additional questions. Tables 3 and 4 present the results of these three models.

In Model 2, response times for questions where the respondent refused to provide an answer were significantly faster than substantive answers. The magnitude of this increase in speed is the among the largest of any of the fixed effects from any of the models. One of our main questions of interest, whether location of the interview impacted response times, was not significant in Model 2, that is, irrespective of respondent characteristics. However, in Model 3, which includes fixed effects for respondent characteristics, conducting interviews in the office was associated with significantly slower response times relative to in respondents' homes, while conducting interviews at a neutral location showed no significant effect. Two respondent characteristics were also significantly faster: female and employed respondents. Congruent with previous literature, respondents with lower education levels and those who were 50 years of age or older had slower response times. Also with slower response times were interviews conducted in a language other than English and respondents who spoke a language other than English, regardless of interview language.

Model 4 introduced item characteristics in addition to location, answer substantiveness, and respondent characteristics. In line with previous research and intuition, both longer questions with more words and more complex questions with more characters per word had significantly slower response times. Questions that started a new section with introductory text, even after accounting for the extra words, also had significantly slower response times. Where questions were repeated multiple times with only the fill for a person's name changed, second iterations had faster response time, with third and further iterations faster still. Questions located in a battery with common answer choices and sometimes sharing a question stem or introductory statement diverged depending on their location. The first question in a battery showed slower response times, while the second and following questions had faster response times, at about the same magnitude of the first question, but in the opposite direction. Broadening the lens to question placement in the interview a whole, questions located later had faster response times.

Compared to questions asking for open-ended answers, all more limited answer choices had faster response times. Ordinal response categories were the fastest, followed by binary answers and closed numerical scales, e.g., "on a scale from 1-10". Compared to basic demographic questions, questions asking for recalled behaviors has slower response times. Finally, questions that asked respondents to look at a list of names on a card they had generated and identify which had a particular

⁵ There were too few separate interview locations to model as a random effect.

⁶ Based on likelihood ratio tests, the inclusion of interviewer random effects terms did not improve the model ($\chi^2_2 = -29,572.92, p = 1.000$); Model 1 had an AIC of 533553.0 and a BIC of 533596.5 (compared to an AIC of 563129.9 and BIC of 563195.1 for the original five-classification model). The addition of only one random effect term for the main, first chair interviewer also did not improve model fit ($\chi^2_1 = -25,093.73, p = 1.000$). The addition of random effects for respondents and items, however, did significantly improve the fit of the model over a linear regression model and multilevel models including either respondents or items ($p < .0001$ for all three likelihood-ratio tests).

Table 3: Cross-classified multi-level model coefficients predicting log(response time in seconds)

	Model 2			Model 3			Model 4		
	Coeff	SE		Coeff	SE		Coeff	SE	
Location (reference: In-home)									
In-office	0.022	0.015		0.030	0.013	*	0.033	0.013	*
Other location	0.000	0.018		0.023	0.015		0.022	0.015	
Respondent characteristics									
Education, less than high school				0.093	0.020	***	0.090	0.020	***
Age 50 or older				0.120	0.011	***	0.120	0.011	***
Female				-0.025	0.010	*	-0.026	0.010	**
Employed				-0.024	0.010	*	-0.024	0.010	*
Interview in a language other than English				0.206	0.019	***	0.202	0.019	***
Speaks a language other than English				0.057	0.009	***	0.056	0.009	***
Question and answer characteristics									
Number of words in question (centered)							0.026	0.002	***
Average number of characters per word (centered)							0.068	0.026	*
Chapter introduction							0.416	0.010	***
Optional question text ("if necessary")							-0.018	0.049	
Interviewer instructions provided							-0.048	0.079	
Follow-up question about person named by R							-0.032	0.058	
Looped question (reference: not iterated / first iteration)									
Second iteration							-0.222	0.004	***
Third or later iteration							-0.300	0.004	***
Battery items (reference: non-battery question)									
First question							0.395	0.096	***
Second or higher question							-0.329	0.059	***
Question order (hundreds) (centered)							-0.041	0.004	***
Binary (yes/no or true/false) answers							-0.562	0.086	***
Closed ordinal answers							-0.859	0.352	*
Closed nominal answer choices							-0.547	0.165	**
Open-ended numerical answer choices							-0.106	0.087	
Select all that apply answer choices							0.040	0.121	
Number of response options							0.036	0.025	
Scale with all choices labeled							0.353	0.301	
Scale with only end points labeled							0.264	0.311	
Type of question (reference: demographic)									
Behavior							0.215	0.067	**
Factual							0.098	0.074	
Opinion							0.137	0.086	
List of people R named are reponse options							0.690	0.162	***
Response card with answer choices							0.204	0.124	^
Refused	-0.785	0.017	***	-0.789	0.017	***	-0.773	0.017	***
Don't know	-0.017	0.011		-0.019	0.011	^	-0.009	0.011	
Controls									
Two people in household				0.001	0.013		0.002	0.013	
Three or more people in household				-0.003	0.013		0.012	0.014	
Child in household				-0.013	0.014		-0.009	0.014	
Caregiver interview	-0.010	0.015		0.019	0.015		0.010	0.015	
Constant	2.243	0.029	***	2.191	0.032	***	2.393	0.094	***

*** $p < .001$, ** $p < .01$, * $p < .05$, ^ $p < .10$

characteristic (e.g., “Who was born in the United States?”) had slower response times, while questions involving other interview aids such as response cards with answer choices printed on them had marginally slower response times.

Table 4: Cross-classified multi-level random effects and model statistics

	Model 1		Model 2		Model 3		Model 4	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Random effects								
Item variance	0.281	0.021	0.280	0.021	0.280	0.021	0.081	0.006
ICC (%)	51.7%		51.6%		52.6%		24.9%	
Respondent variance	0.036	0.001	0.036	0.001	0.026	0.001	0.027	0.001
ICC (%)	6.6%		6.7%		4.9%		8.1%	
Residuals	0.227	0.001	0.226	0.001	0.226	0.001	0.220	0.001
ICC (%)	41.7%		41.6%		42.4%		67.0%	
Model Fit								
Log likelihood	-266772.5		-265712.4		-265471.1		-259687.7	
AIC	533553.0		531442.8		530978.2		519459.4	
BIC	533596.5		531540.6		531173.8		519915.8	
<i>df</i>	4		9		18		42	
<i>n</i>	387,434		387,434		387,434		387,434	

Somewhat surprisingly, neither the number of answer choices nor an item allowing respondents to select all that apply rather than forcing one choice had significantly different response times. The labelling on numeric scales did not significantly affect response times, nor did asking for a number rather than any form of open-ended answer. Questions that asked for opinions or other types of facts than demographics were similarly not significantly faster or slower, nor were optional text for interviewers to guide respondents if necessary or additional instructions directed only to interviewers.

5. Discussion and Conclusion

In this paper, we presented analyses exploring the joint associations of interview location, interviewers, respondent characteristics, and item—both question and answer—characteristics with, and their relative contributions to variability in, survey response times in an in-person interview of low-income New York City applicants to affordable housing. Our findings add to a small but growing literature on survey item response times that leverages the automated collection of item-level time stamps by CAI software. Beyond theoretical interest in the cognitive processes underlying survey response, the relevance of response times to survey research is generally conceptualized either in terms of the efficiency of data collection—including burden on respondents and costs to the sponsor—or for data quality analysis as a proxy for satisficing or processing problems (Couper & Kreuter, 2013; Dahlhamer et al., 2020; Garbarski et al., 2020; Olson & Smyth, 2015; Sturgis et al., 2021). Practically, shorter response times per question can translate directly to either more questions per interview or fewer resources expended per case for the same number of questions.⁷ Adopting a view of the face-to-face survey interview as fundamentally a social interaction—a conversation between interviewer and respondent—we expand our assessment of the determinants of response times to include social and cultural factors and contemplate the potential for a novel view of response times as tool for indexing the respondent survey experience.

⁷ For example, a recent message sent to AAPORnet (a members-only listserv for survey researchers) offered space to add custom questions to an online Omnibus survey. Items were priced at \$40-per-second, providing the example of a 10-12 second question costing \$400-\$480.

Aiming to bolster the evidence base on response times, our analysis first sought to replicate the findings of others, with many of the findings presented in this paper reinforcing similar analyses done on other surveys. Respondent factors that were associated with slower response times included interviews that were conducted in languages other than English, lower levels of education, and older age; slower item characteristics included more words, more characters per word, the first item of a battery, and question types other than demographics. Respondent factors associated with faster response times included employment; item characteristics included later position in a battery and answer choices other than open-ended. Other findings diverged from the existing body of research. According with Sturgis et al. (2021), we found that female respondents had faster response times, while Dahlhamer et al. (2020) reported the opposite and Olson and Smyth (2015) found no association. We found that respondents who spoke a language other than English had slower times, while Dahlhamer et al. (2020) found no association between response times and nativity. The sociocultural nature of these two factors may explain the lack of consistent results throughout the literature. We found item refusal was associated with faster response times, while Draisma and Dijkstra (2004) found that non-substantive answer took longer to provide. A respondent's willingness to provide answers—and, conversely, comfort in openly refusing to answer rather than demurring by pleading ignorance—are partly products of the rapport with the interview in the social interaction. Finally, against a background of mixed evidence, we found that later placement in the instrument was associated with faster response times, suggesting that item location may be sensitive to social factors such as waxing or waning rapport between interviewer and respondent or situational context within a given study.

The model specifications used in this analysis assume that the respondent characteristic effects are constant over items and the effects of item features are constant across respondents. Yan and Tourangeau (2008) relaxed these assumptions and allowed the effects of age to vary randomly over items and the effects of question types, types of answer scales, and answer labels to vary randomly over respondents. They found that the association between older respondents and slower response times did vary significantly across items, and question and answer associations varied across respondents. However, they did not test which types of questions were faster or slower by age or which respondent characteristics predicted faster or slower response times for which question features. Future research should continue in this vein to relax the assumption of constant effects by not only incorporating random-effects of specific characteristics but also modeling specific respondent-item interactions to identify which types of respondents answer which types of questions faster or slower. Though previous analyses have used this modeling strategy to examine the interaction between interviewer and respondent or interviewer and item characteristics (Olson & Smyth, 2015; Sturgis et al., 2021), to the best of our knowledge this has not been done for item and respondent characteristics. This information would enable survey designers to tailor their questions to their population of interest, or at least be aware of potential pitfalls.

Consistent with our broader conception of the determinants of response times including sociocultural factors, this analysis examined some novel dimensions of interview setting and specific item features that were found to be significantly associated with response times. Interviews conducted at the project offices were associated with slower response times compared to those in respondents' homes when including respondent characteristics in the model. Interviews in neutral, third locations, however, were not significantly different. To our knowledge, no study has explored the impact on interview location on item response times. Respondents may have answered more slowly in the office because of fewer distractions than at home. Some may have answered more cautiously because of discomfort in a government building and associated fears of surveillance or the heightened power dynamic in that setting. Future research should examine variation in data quality by location to assess these different explanations of this response time finding. However, given that location was only significantly associated with response times when respondent characteristics were controlled for and that respondents were not randomly assigned to interview location, this location "effect" may be confounded by respondent self-selection into in-office interviews.

Questions preceded by an introduction to a section were associated with slower response times of a large magnitude, even after controlling for the word count. This may be because switching between topics required additional cognitive processing by respondents (Tourangeau et al., 2000), but this finding may also be rooted in conversational norms around changing topics. The response times for questions when iterated a second and following time across people—such as asking for demographics of household members one at time—were associated with faster response times. This likely represents respondent learning the pattern of the questions, and potentially anticipating a loop and even providing an answer before the question was completely read. NYCHANS used a variety of physical aids to support the interview experience. A unique contribution of this paper is testing the association with response times of another way of collecting the sort of information generally collected in iterated questions—adopting a reference card with a list of people’s names, such as a roster card. We analyze two types of questions that relied on this card. Questions where the reference card served as a list of potential responses (“Who on this card is female?”) were associated with slower response times compared to all other questions. This was likely due to the complexity of the cognitive process of assessing each individual for the trait asked about.

But we did not assess the relative total speed of asking a single select-all-that-apply question with potential responses listed on the card compared to a series of identical yes/no question over each person separately. Emphasizing that design decisions should primarily be based on the planned analysis, Olson and Smyth (2015) include the format of response options in their classification of “necessary question features” as those essentially required by the purpose of a survey that narrowly constrain the researchers crafting the questionnaire. But this iteration/select-all-that-apply choice is one example of many where multiple ways of constructing items to measure the same construct compete without a clear way to distinguish the optimal format. Many would argue that the measurement properties of each item should determine which is preferred to reduce total error. In this example, forced choice yes/no items have been shown generally to have higher rates of endorsement and corresponding signs of optimizing rather than acquiescence bias (Stenger et al., 2023).⁸ But the introduction of interview aids to support respondent optimizing in select-all-that-apply questions such as cards may moot the controversy. In general, though, for flexible equivalent situations like these—which are more numerous than is often recognized (Harder, 2020)—it is useful to understand total response times for each method. Perhaps choosing the faster or slower version of the questions for that location would benefit the balance and tempo of the questionnaire overall. Beyond speed alone, analyses that focus on the contribution to the holistic respondent experience of these competing variants for collecting basic information about many people would be particularly valuable.

The most consistent finding throughout the response times literature relates to the relative contribution of interviewers, respondents, and individual items to variation in response times. All studies that have utilized a cross-classified multilevel approach to examine all three have found that interviewers contribute the smallest amount, respondents contribute a larger but still relatively small share, and items contribute the largest share. The findings of the analysis presented here were consistent with this pattern. Despite their small share—generally 3% or less—interviewer contributions to response times often draw special attention as a uniquely valuable method to isolate and examine interviewer effects. In this analysis, this interviewer contribution was insubstantial to the point that models that did not account for interviewers fit the data better than a model structure that included the interviewers.⁹ But, as Sturgis et al. (2021) cautioned, in larger data collection

⁸ Incorporating response times into an analysis of this sort could provide evidence that slower response times for these items is associated with optimizing and is beneficial.

⁹ A formally similar analysis by Olson, Smyth, and Ganshert (2019) used multilevel cross-classified models to decompose relative interviewer, respondent, and item contributions to variance in respondent behaviors (rather than response times) and found essentially zero variability across interviewers in respondent answering, with proportions ranged from 0% for answers that the respondent qualified in some way to 2.5% for “don’t know” responses and refusals combined and 3.2% for “don’t know” responses alone. Though the study did not examine response times alongside respondent behaviors, these results suggest that our null findings for interviewer effects on response times may not be unique and that interview contributions to response time variation may not translate into other measurable differences, including in data quality.

agencies with large pools of interviewers and limited ability to carefully oversee individual interviewers, there may be much more variability. Higher shares of interviewer contribution to response time variance are particularly likely where interviewers are hired as part-time workers specifically for a short-term data collection effort with little or no previous experience (for example, Dahlhamer et al. (2020) showed that US Census Bureau interviewers contributed over 9% of response time variance in the 2014 NHIS). Consistent across these analyses is that items are the largest contributors to response time variance with respondent characteristics in a distant second place. This holds promise for survey designers, who directly control the largest source of variation—the items (Olson & Smyth, 2015).

But these relative contributions to the variance provide specific information about the relative hetero- and homogeneousness of the response times between interviewers, respondents, and items in a given survey and its study population. In non-technical terms, the item contribution, for example, tells us how consistently similar or different response times are across different items. If items in an interview are similar and have basically similar response times, the between-item contribution to overall variance would be negligible. If, on the other hand, some items are short and fast, while others are long and slow, items would account for a larger share of the variance. Ultimately, the variation attributed to each level or classification of the model in each analysis—interviewers, respondents, items—tells us about the similarity or difference within each level or classification in a particular study. Most research has been conducted on relatively long questionnaires because they provide more items and response times to analyze, and as questionnaires get larger, items are naturally more varied in type, topic, length, and other features. Because the variation partition is wholly contingent on the heterogeneity of items, respondents, and interviewers, respectively, we should not be surprised that the variation in response times attributable to items in the literature has generally been large. More response time research should be conducted on shorter questionnaires—or at least those with relatively similar items—administered to respondents who exhibit diversity in the characteristics that are associated with response times.

Some survey designers would argue that, all things equal, faster responses are better. Others would advocate for an optimal time for each type of question—neither too fast nor too slow—that ensures the highest data quality in the shortest time possible. When researchers fielding surveys plan around respondent burden, the main consideration is usually total time expected of respondents across the interview interaction. Within the instrument itself, survey designers are more likely to consider varying item modes, topics, and types, balancing cognitive and emotional intensity, and aligning response options to support data processing and analysis. But in reflecting on item-level response times, it seems that their collective progression can be conceptualized as a measure of the respondent survey experience. As a social interaction—a conversation with a purpose—an interview has a life to it, a cadence and rhythm and tempo. Response times could be included as a component in designing data collection around the respondent experience. The fact that items usually contribute the largest share of variation in response times indicates that most questionnaires already use questions of widely varying length and likely already vary tempo throughout. But with rhythm as an explicit consideration, the ordering of questions could be more precise when guided by response times and their determinant item characteristics.¹⁰ Deliberately varying item lengths throughout an interview may improve and modulate that experience throughout, contributing to increased data quality, fewer breakoffs, and lower item nonresponse, among other benefits.

Acknowledgements

The authors wish to acknowledge the hard work and dedication of both first and second chair interviewers who worked tirelessly and painstakingly on NYCHANS to reduce both their effects on survey error and their contributions to variation in response times as well as the respondents who

¹⁰ We are not suggesting that response times be the only criteria for item order. Rather, they should be considered, along with other documented context and order effects, which can arise in unexpected items (see, for example, Goldstein and Gaumer (2023)).

gave generously of their time and effort. We apologize for any shortcomings in the questions and answers they were requested to ask and answer and are grateful for the lessons they taught us, directly and indirectly, about creating good survey items and crafting effective questionnaires. Without them this work would not be possible and there would be no papers to write.

References

- Andreadis, I. (2021). Web survey response times: What to do and what not to do. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1774–1782. <http://www.asasrms.org/Proceedings/y2021/files/1912254.pdf>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Bassili, J. N., & Krosnick, J. A. (2000). Do strength-related attitude properties determine susceptibility to response effects? New evidence from response latency, attitude extremity, and aggregate indices. *Political Psychology*, 21(1), 107–132. <https://doi.org/10.1111/0162-895X.00179>
- Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, 60(3), 390. <https://doi.org/10.1086/297760>
- Bell, K., Fahmy, E., & Gordon, D. (2016). Quantitative conversations: The importance of developing rapport in standardised interviewing. *Quality & Quantity*, 50(1), 193–212. <https://doi.org/10.1007/s11135-014-0144-2>
- Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 313–334). Routledge/Taylor & Francis Group.
- Bergmann, M., & Bristle, J. (2020). Reading fast, reading slow: The effect of interviewers' speed in reading introductory texts on response behavior. *Journal of Survey Statistics and Methodology*, 8(2), 325–351. <https://doi.org/10.1093/jssam/smy027>
- Bingham, W. V. D., & Moore, B. V. (1959). *How to interview* (4th rev. ed). Harper. (Original work published 1934)
- Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1(2), 103–124. <https://doi.org/10.1177/1471082X0100100202>
- Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *BMS: Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 103(1), 5–25. <https://doi.org/10.1177/075910630910300103>
- Campanelli, P., & O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey nonresponse. *Quality and Quantity*, 33(1), 59–76. <https://doi.org/10.1023/A:1004357711258>
- Couper, M. P. (1996). Changes in interview setting under CAPI. *Journal of Official Statistics*, 12(3), 301–316.
- Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176(1), 271–286. <https://doi.org/10.1111/j.1467-985X.2012.01041.x>
- Dahlhamer, J. M., Maitland, A., Ridolfo, H., Allen, A., & Brooks, D. (2020). Exploring the associations between question characteristics, respondent characteristics, interviewer performance measures, and survey data quality. In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation, and testing* (pp. 153–192). International Conference on Questionnaire Design, Development, Evaluation, and Testing, Hoboken, NJ. Wiley.
- Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17(1), 30–52. <https://doi.org/10.1177/1525822X04269550>
- Draisma, A. R., & Dijkstra, W. (2004). Response latencies and (para)linguistic expressions as indicators of response error. In S. Presser (Ed.), *Methods for testing and evaluating survey questionnaires* (pp. 131–147). John Wiley & Sons, Inc.

- Garbarski, D., Dykema, J., Schaeffer, N. C., & Farrar Edwards, D. (2020). Response times as an indicator of data quality associations with question, interviewer, and respondent characteristics in a health survey of diverse respondents. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 253–266). CRC Press, Taylor & Francis Group.
- Goldstein, D., & Gaumer, E. (2022). Making lotteries legible: Designing natural experiments. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association.
http://www.asasrms.org/Proceedings/y2022/files/401613_502926.pdf
- Goldstein, D., & Gaumer, E. (2023). How many (bed)rooms? Hard to detect question order effects in factual questions. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. <https://doi.org/10.5281/ZENODO.10002318>
- Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5), 1158–1177.
<https://doi.org/10.1177/1745691620917678>
- Heerwegh, D. (2002). *Describing response behavior in websurveys using client side paradata*. [Paper presentation]. International Workshop on Web Surveys. Mannheim, Germany.
- Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, 21(3), 360–373.
<https://doi.org/10.1177/0894439303253985>
- Herzog, H. (2012). Interview location and its social meaning. In J. Gubrium, J. Holstein, A. Marvasti, & K. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (pp. 207–218). SAGE Publications, Inc.
<https://doi.org/10.4135/9781452218403.n14>
- Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74(5), 956–984.
<https://doi.org/10.1093/poq/nfq062>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236.
<https://doi.org/10.1002/acp.2350050305>
- Leverentz, A. (2023). Interview location as data. *Qualitative Sociology*, 46(4), 489–514.
<https://doi.org/10.1007/s11133-023-09548-4>
- Loewen, E. R., Bauer, E., Thompson, M. E., Martin, N., Quah, A. C. K., & Fong, G. T. (2022). Timing estimates for complex programmed surveys. *Survey Practice*.
<https://doi.org/10.29115/SP-2022-0011>
- Loosveldt, G., & Beullens, K. (2013). The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research*, 42(6), 1422–1430.
<https://doi.org/10.1016/j.ssresearch.2013.06.005>
- Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, 72(5), 914–934. <https://doi.org/10.1093/poq/nfn050>
- Miller, K. (2004). Implications of socio-cultural factors in the question response process. In P. Prüfer, M. Rexroth, & F. J. J. Fowler (Eds.), *QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards, 21–23 October 2003* (pp. 172–189). Zentrum für Umfragen, Methoden und Analysen -ZUMA-.
<https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49209-8>
- Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. Times Books Henry Holt and Company.
- Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata* (1st ed., pp. 43–72). Wiley.
<https://doi.org/10.1002/9781118596869.ch3>
- Olson, K., & Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*, 3(3), 361–396.
<https://doi.org/10.1093/jssam/smv021>
- Olson, K., Smyth, J. D., & Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, 7(2), 275–308. <https://doi.org/10.1093/jssam/smy006>

- Pirralha, A., Haag, C., & Maurice, J. v. (2023). Adjusting to the survey: How interviewer experience relates to interview duration. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, 18(2), 185-212. <https://doi.org/10.12758/MDA.2023.05>
- Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using Stata* (Fourth edition). Stata Press Publication.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510–532. <https://doi.org/10.1037/0033-2909.114.3.510>
- Schaeffer, N. C. (2004). Conversation with a purpose—or conversation? Interaction in the standardized interview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (2nd ed., pp. 365–391). Wiley. <https://doi.org/10.1002/9781118150382.ch19>
- Schaeffer, N. C., Dykema, J., & Maynard, D. W. (2010). Interviewers and interviewing. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 437–471).
- Schneider, S., Jin, H., Orriens, B., Junghaenel, D. U., Kapteyn, A., Meijer, E., & Stone, A. A. (2023). Using attributes of survey items to predict response times may benefit survey research. *Field Methods*, 35(2), 87–99. <https://doi.org/10.1177/1525822X221100904>
- Schouten, B., van den Brakel, J., Buelens, B., Giesen, D., Luiten, A., & Meertens, V. (2021). *Mixed-mode official surveys: Design and analysis* (1st ed.). Chapman & Hall/CRC. <https://doi.org/10.1201/9780429461156>
- Stenger, R., Olson, K., & Smyth, J. (2023). *Meta-analysis of multiple response survey questions: Question characteristics that influence cognitive processing in check-all and forced-choice formats*. [Conference presentation]. AAPOR 78th Annual Conference, Philadelphia, PA. <https://aapor.confex.com/aapor/2023/meetingapp.cgi/Paper/1229>
- Sturgis, P., Maslovskaya, O., Durrant, G., & Brunton-Smith, I. (2021). The interviewer contribution to variability in response times in face-to-face interview surveys. *Journal of Survey Statistics and Methodology*, 9(4), 701–721. <https://doi.org/10.1093/jssam/smaa009>
- The American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.). AAPOR.
- Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2), 169–181. <https://doi.org/10.1108/QAE-06-2017-0034>
- Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68(3), 368–393. <https://doi.org/10.1093/poq/nfh035>
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Turner, G., Sturgis, P., & Martin, D. (2015). Can response latencies be used to detect survey satisficing on cognitively demanding questions? *Journal of Survey Statistics and Methodology*, 3(1), 89-108. <https://doi.org/10.1093/jssam/smu022>
- Waickman, C. R., Goldstein, D., Powell, L. M., & Gaumer, E. (2019). Choreographing “the best interview ever”: Developing and implementing a multimodal family interview. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1735–1747. <http://www.asasrms.org/Proceedings/y2019/files/1199571.pdf>
- Wenz, A. (2015). *Predicting response times in web surveys*. [Conference presentation]. 17th General Online Research Conference, Cologne, Germany.
- West, B. T., & Blom, A. G. (2016). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175-211. <https://doi.org/10.1093/jssam/smw024>
- Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (1st ed., pp. 73–95). Wiley. <https://doi.org/10.1002/9781118596869.ch4>
- Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1), 51–68. <https://doi.org/10.1002/acp.1331>
- Yu, E., Terry, R. L., Kline, A., Fee, H., & Kaplan, R. (2019, February 26). *Exploring the impact of interviewer perceptions and interviewer-respondent interactions on the survey of income and program participation: analysis of CARI recordings*. [Paper presentation]. Interviewer

Workshop, 2019: Interviewers and Their Effects from a Total Survey Error Perspective.
<https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=sociw>

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2), 127-135.
<https://doi.org/10.18148/SRM/2014.V8I2.5453>