9-2020

The effect of emphasis in telephone survey questions on survey measurement quality

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

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

The effect of emphasis in telephone survey questions on survey measurement quality

Kristen Olson and Jolene D. Smyth

Department of Sociology, University of Nebraska-Lincoln,
703 Oldfather Hall, Lincoln, NE, 68588-0324, USA

Correspondence — Kristen Olson kolson5@unl.edu Department of Sociology,
University of Nebraska–Lincoln, , Lincoln, NE 68588-0324, USA

Abstract

Questionnaire design texts commonly recommend emphasizing important words, including capitalization or underlining, to promote their processing by the respondent. In self-administered surveys, respondents can see the emphasis, but in an interviewer-administered survey, emphasis has to be communicated to respondents through audible signals. We report the results of experiments in two US telephone surveys in which telephone survey questions were presented to interviewers either with or without emphasis. We examine whether emphasis changes substantive answers to survey questions, whether interviewers actually engage in verbal emphasis behaviors, and whether emphasis changes the interviewer- respondent interaction. We find surprisingly little effect of the question emphasis on any outcome, with the primary effects on vocal intonation and the interviewer-respondent interaction. Thus, there is no evidence here to suggest that questionnaire designers should use emphasis in interviewer-administered questionnaires to improve data quality. As the first study on this topic, we suggest many opportunities for future research.

Keywords: Telephone surveys, questionnaire design, interviewer-respondent interaction, survey methodology

Published in *International Journal of Social Research Methodology* (2020)
doi:10.1080/13645579.2020.1824628

Published 28 September 2020.

Introduction

Questionnaire designers often use emphasis in survey questions to indicate words that are particularly important for the meaning and understanding of the question (Dillman et al., 2014, p. 174; Redline, 2011; Sudman & Bradburn, 1982, p. 238) such as a time frame, variation in response categories, or instructions. Emphasis may be conveyed using font (Times New Roman, Calibri), case (all caps), and font style (bold, italic, underlined). For instance, the National Science Foundation's self-administered Survey of Earned Doctorates uses a mix of capital letters, underlining, and bold to emphasize words (e.g., 'A6. Which TWO sources listed in A5 provided the most support? Enter **letters** of primary and secondary sources' (National Center for Science and Engineering Statistics [NCSES], 2016)). The interviewer-administered U.S. Current Population Survey uses capital letters ('LAST WEEK, did you do ANY work for pay?' (Bureau of Labor Statistics [BLS], 2014)). The US National Center for Education Statistics Early Childhood Longitudinal Study: Kindergarten Class of 2010–2011 interviewer-administered Fall Parent Interview uses underline to indicate emphasis in the question text: 'Has {CHILD} ever received care from a relative on a regular basis?' (National Center for Education Statistics [NCES], 2012). Despite widespread use, few systematic evaluations of emphasis as a questionnaire design tool have been carried out (Falcone, 2017; Falcone et al., 2018; Redline, 2011, 2013), especially in interviewer-administered surveys.

Respondents can see emphasis used in self-administered surveys. In interviewer-administered surveys, however, the researcher must convey the question's emphasis to interviewers who may then use pitch and tone of voice to communicate this emphasis to respondents. To our knowledge, no published studies examine whether and how interviewers read emphasized words or the effect emphasis has on interviewer/respondent behaviors and answers to survey questions. This paper does so, using two national telephone surveys. It aims to answer three research questions:

RQ1: Does emphasizing words in survey questions influence interviewers' pitch and intonation?
RQ2: Does emphasizing words affect interviewer/respondent behaviors during the interview?
RQ3: Does emphasizing words affect respondents' answers?

Literature review

Readers infer meaning from changes in font, style, and case of written words (Doyle & Bottomley, 2009). In particular, emphasized written words may lead to more visual processing, awareness of, and retention of emphasized words (Fraundorf, 2012; Fraundorf et al., 2010; Margolin, 2013; McAteer, 1992; McCarthy & Mothersbaugh, 2002; Sanford et al., 2006). Sanford et al. (2006) found, for example, that participants were better able to notice and report changes in text when the changes were italicized, suggesting they more deeply process italicized text. Findings like these are reflected in the conventional wisdom (e.g., Dillman et al., 2014; Redline, 2011; Tourangeau et al., 2013) that important words in survey questionnaires should be emphasized.

In self-administered surveys, it is believed that such emphasis will draw respondents' attention, impressing upon them the importance of the emphasized material and helping them interpret the material as desired by the researcher (Crawford et al., 2005; Dillman et al., 2014; Redline, 2011; Tourangeau et al., 2013). As a result, some researchers have adopted the practice of emphasizing elements they feel are essential for interpretation like time frames (e.g., 'in the last year'), inclusion/ exclusion instructions (e.g., 'including yourself'), and the referent of the question (e.g., you, your spouse, etc.).[1] In contrast with this perspective, Dillman et al. (2014) suggest that changing the font of instructions (e.g., italicizing them) can signal that the instructions may be optional or not needed by all respondents, thus causing them to pay less attention to the content.

Few studies we know of examine emphasis in questionnaire design. Redline (2011, 2013) experimentally assigned respondents to receive a web survey with instructions in the same font as the question stem or different (i.e., italicized) and found no significant effect of the emphasis on responses. Similarly, in a survey app, Falcone (2017) found no differences in responses between treatments with instructions in normal versus italic font, but found that the italicized version was completed significantly quicker. In a follow-up eye tracking study, Falcone et al. (2018) found that respondents fixated fewer times and for less time on instructions when they were italicized. These results suggest that visual emphasis in self-administered surveys may influence respondent reading behaviors, but seem to have little effect on answers. However, we know of no studies that have tested the impact of emphasis on important words

or phrases within a question stem (i.e., outside of instructions) in either self- or interviewer-administered surveys.

Outside of survey methodology, research on emphasis has shown that speakers communicate emphasis through changes in pitch, intonation, duration of speaking, and volume (Fraundorf et al., 2010; Lieberman, 1960; Ryalls & Behrens, 2000; Sanford et al., 2006) with pitch being the most common acoustic component used for word stress (Lieberman, 1960). In a sentence with both emphasized words and non-emphasized words, emphasized words are spoken at a higher pitch and words that are not emphasized are spoken at a decreased pitch (Cooper et al., 1985), except for the first word in a sentence, which typically already has a higher pitch due to location so instead tends to be emphasized by increased duration. In addition, when a focal word is emphasized versus not emphasized, the pitch on the next word in the sentence after the focal word drops (Cooper et al., 1985). Thus, the average pitch across a sentence is likely to be similar, but there will be more variation in pitch, that is, greater intonation, when words are emphasized. Although much of this research is conducted with the text of a quite different nature from surveys (e.g., short declarative sentences; sentences in a story) and often examines variation within a person over different sentences, in absence of similar research within the survey context, we use it to generate our hypotheses. Thus, *we hypothesize that questions will be read with similar mean pitch (H1a) and more intonation (variation in pitch) (H1b) when they are visually emphasized compared to when they are not visually emphasized*.

Increased inflection or intonation in spoken words changes how the listener processes the material (Ito & Speer, 2006), cueing the listener that the emphasized word is new and thus requires more attention and increasing the depth of listener processing (Birch & Clifton, 2002; Sanford et al., 2006). For example, Sanford et al. (2006) found that listeners were more likely to identify subtle changes in text that was read to them when the changed word was emphasized with a change in pitch. In addition to drawing attention to specific words to help listeners identify changes, emphasis can be used to alter how listeners interpret identically worded sentences. For example, Carlson et al. (2009) found that how listeners interpreted the meaning of a vague phrase depended on which words in a sentence containing the phrase were emphasized with pitch.

Thus, emphasis can draw the attention of the reader to important information *and* influence meaning for listeners in strategic ways. In questionnaires, it should cue the interviewer to take care in reading and cue respondents that certain information in the question is particularly important. As a result, one hypothesis is that questions containing emphasis will be *more* likely than those that do not to yield paradigmatic sequences (H2a). In a paradigmatic sequence, the interviewer asks the question as written, the respondent provides an adequate and codable answer, and the interviewer may or may not provide a brief acknowledgment (Schaeffer & Maynard, 1996). Any deviations from these behaviors by either actor such as giving feedback, asking for clarification, or probing is nonparadigmatic.

Alternatively, emphasis may cue both interviewers and respondents to be more vigilant about the accuracy of responses, which may lead them to engage in additional conversation in an effort to get it right. For example, it may make interviewers more likely to repeat emphasized words, provide clarification, probe ambiguous answers or verify answers, and it may trigger respondents to ask for clarification of the meaning of emphasized words or to use the emphasized words to answer, qualify, or give context to their answers. This scenario yields the competing hypothesis that questions with emphasis will be less likely than those without to yield paradigmatic sequences (H2b).

Additional conversation between interviewers and respondents takes time (Timbrook et al., 2018). Thus, if emphasis increases the probability of paradigmatic sequences, we hypothesize that it will also decrease the total amount of time spent on a question, including question asking, question answering, and any necessary follow-up and feedback behaviors (which we will call response time, following Olson & Smyth, 2015) (H2c). However, if emphasis decreases the probability of paradigmatic sequences, we hypothesize that it will also increase response time (H2d).

What about the effect of emphasis on the actual responses given to the survey question? First, if having emphasis, compared to not having emphasis, more effectively communicates important information that cues interviewers to the importance of the question or helps respondents cognitively process the important parts of a survey question, then it should reduce item nonresponse rates. Thus, we hypothesize that questions with emphasis will have lower item nonresponse rates than the same questions without emphasis (H3a). Conversely, if emphasis

cues the respondent to the question being especially important, it is possible that the respondent will be more likely to respond 'don't know' than to provide an answer about which they may be less certain (H3b).

The impact of emphasis on responses should be highly dependent on which words are emphasized in a question and how the emphasis clarifies the meaning of the question. As such, we need to introduce the types of questions for which emphasis is evaluated in this study. **Table 1** shows the question wording and response options for the experimental items included in this paper, which come from two surveys, the Work and Leisure Today 2 (WLT2) survey and the Consumer Spending (CS) survey. The emphasis in the first set of questions is on the time frame respondents should consider when answering. In general, we expect higher reports for items with long time frames (e.g., ever, H3c1, H3c2) and lower reports for items with shorter time frames (e.g., the last year, monthly, H3c3, H3c4) simply because a longer time frame provides more opportunity for an event to occur.[2] Inasmuch as the emphasis makes the time frame more obvious and thus increases the likelihood of respondents abiding by it, we expect stronger effects (see Table 5 for summary of expectations by reference period) when emphasis is used than when it is not used.

The next type of emphasis in Table 1 clarifies the referent of the question; that is, whether the question is about 'you' or 'anyone else.' When the referent 'you' is weaker, as in the no emphasis version, we expect higher reports as respondents may mistakenly include others in the household in their report. However, when the referent 'you' is stronger, as in the emphasized version, we expect higher compliance (i.e., respondents answer only for themselves) and thus lower reports (H3d1). The same logic applies to the item referring to 'anyone else,' but the direction is opposite. When 'anyone else' is emphasized, we expect higher reports, especially in households with more members (H3d2).

The third type of emphasis in Table 1 clarifies who should be included/excluded in reports. All three questions tested here have instructions to 'include yourself' in counts of household members. We expect this instruction to be more obvious and thus more likely to be followed in the emphasis condition, thus increasing reports compared to the no emphasis condition (H3e). Finally, the emphasis in the last two items is on response options. Previous research on vocal emphasis in laboratory research shows that experimenters vocally emphasizing one

Table 1. Question wording and response distribution hypotheses for emphasis experiment by type.

Survey	*Question stem*	*Response options*
TIME FRAME		
WLT2	Q2. Compared to 10 YEARS AGO IN 2005, do you think people have more leisure time, less leisure time or about the same amount?	More, Same amount, Less
WLT2	Q5. Have you EVER been employed for pay or profit?	Yes, No
WLT2	Q6. Have you EVER been laid off from a job?	
WLT2	Q7. Have you EVER been fired from a job?	
CS	Q11. Have you EVER participated in a sharing economy service as either an OWNER or RENTER?	
WLT2	Q32. During THE LAST YEAR, how many parking tickets have you received?	Enter number
WLT2	Q33. During THE LAST YEAR, how many speeding tickets have you received?	
CS	D45. What is your total MONTHLY household income, before taxes?	Ranges from Under $60 to $20,000 and over
CS	D46. Is your total MONTHLY household income before taxes $4000 or more, or is it less than $4000?	DK/REF lead to D46
REFERENT OF QUESTION		
WLT2	Q24. The next questions are about mobile phones. For these questions, a cell phone is a mobile telephone on which only calls or texts are made and received. A smartphone is a mobile telephone on which the user can access the internet, use apps, and read email, as well as send and receive calls and texts. Do YOU happen to have a working CELL PHONE OR A SMARTPHONE?	Yes, cell phone; Yes, smartphone; Yes, both a cell phone and a smart phone; No
WLT2	Q25. Does ANYONE ELSE in your household have a working mobile phone, either cell phone or smartphone?	Yes; No
INCLUSION/EXCLUSION		
WLT2	Q42. How many people, INCLUDING YOURSELF, live in your household?	Enter number
WLT2	Q43. How many people, INCLUDING YOURSELF, are adults age 18 and older?	Enter number
CS	D9. INCLUDING YOURSELF, how many adults, 18 years of age or older, live in this household?	Enter number (1–96)
RESPONSE OPTIONS		
CS	Q1. Please think about your total household spending over the past four weeks, and all the purchases you have made. Compared to the same time a year ago, would you say you are spending MORE money, spending LESS money, or about the SAME amount of money?	Spending more; Spending less; Spending about the same
CS	Q2. Compared to a year ago, do you feel MORE willing to spend money on things you may not need, LESS willing to spend money, or do you feel the SAME about spending money on things you may not need?	More willing to spend money; Less willing to spend money; Feel the same about spending money

response option more than another is associated with higher selection of the emphasized response option by subjects (Duncan & Rosenthal, 1968). However, in our experiment, the visual emphasis in the questionnaire is placed equally on each response option. Thus, provided interviewers apply the emphasis equally, we do not have directional hypotheses for responses on these items (H3f).

Data and methods

We used two national random digit dial (RDD) telephone experiments in the United States. First, the Work and Leisure Today 2 (WLT2) survey was designed by the authors and conducted by Abt SRBI during September 2015 using a dual-frame survey (n = 902, Landline = 451, AAPOR RR3 = 9.4%; Cell phone = 451, AAPOR RR3 = 7.1%; The American Association for Public Opinion Research [AAPOR], 2016), with an adult selected using the Rizzo method in the landline frame and the phone answerer selected in the cell frame. The 14-minute long, on average, WLT2 survey asked about work and occupations, leisure activities, technology, and demographics.

In WLT2, sampled phone numbers were randomly assigned to one of two different experimental conditions, emphasis (denoted with all caps) and no emphasis. Each experimental condition was conducted by a separate set of interviewers (n = 451 respondents in each condition; n = 14 interviewers in no emphasis condition, n = 13 interviewers in emphasis condition).[3] Thus, in WLT2, the emphasis assignment is both a property of questions and of interviewers. Interviewers were trained that the words in all caps should be vocally emphasized, although they were provided discretion over how to implement this. This was reinforced during training through 'round robin' reading of the survey questions, in which the emphasized words were read with more vocal inflection. WLT2 interviews were audio recorded and transcribed. They were then behavior coded (described below) and subjected to acoustic measurement using the Sequence Viewer Software (Dijkstra, 2016).

The second study was the Gallup Consumer Spending (CS) survey. This survey was sponsored and fielded by the Gallup Organization during November 2015. CS was also a dual frame national RDD telephone survey; landline respondents were selected using the next birthday

method and the phone answerers were selected as the cell phone respondents. Unlike WLT2, the experimental version of this survey modified a survey already being used by Gallup rather than the questionnaire being fully designed by the authors (Gallup-designed version n = 1517, AAPOR RR1 = 5.6%, RR3 = 8.6%; experimental version, n = 459, AAPOR RR1 = 6.1%, RR3 = 10.8%). The CS questionnaire asked about perceptions of the economy, household spending in different areas, use of peer-to-peer sharing services (e.g., Uber, Airbnb), and demographics.

Because the initial survey was developed by Gallup, emphasis, again denoted by all caps, was added to some questions in the experimental version and removed from other questions. Thus, interviewers were randomly assigned to a questionnaire version containing four questions with emphasis and two questions without emphasis (original Gallup version), or a version containing four questions without emphasis and two questions with emphasis (author version). That is, there was within-survey variation in the use of emphasis in the CS questionnaire. Interviewers were trained to read words with all caps with emphasis, but allowed discretion about how exactly to implement emphasis.

Independent variables

The primary independent variable was whether the respondent answered a question containing emphasis or no emphasis. In each survey, the questions varied in the number of words that were emphasized within the question stem. Because it may be easier to vocally convey emphasis on single words than on multiple words, we also included a variable for the number of words that are emphasized in the question stem as a measure of exposure to the emphasis experiment. In WLT2, 3 questions had one emphasized word (2 questions in CS), 3 questions had two emphasized words (1 question in CS), 3 questions had three emphasized words (3 questions in CS), and 1 question had six emphasized words (no questions in CS). Because the number of emphasized words should only have an effect in the experimental condition where words were actually emphasized, we included an interaction term between whether the respondent was assigned to the emphasis condition and the number of emphasized words.

Dependent variables

We used four types of data for dependent variables – acoustic measures, paradata, behavior codes, and responses. While response data and paradata are common, acoustic measures and behavior codes require extensive, time consuming, and costly data processing; thus, data sets containing these measures, like WLT2, are very rare. Due to resource and time constraints, these measures are not available for CS.

Acoustic measures – WLT2 only

Our first dependent variables were pitch and intonation, which were obtained using Sequence Viewer's Waveform analysis feature (Dijkstra, 2016) to analyze interview audio recordings. All of the interviewers were in the same telephone facility. The audio files for this project were recorded in mono with a sampling rate of 8,000 Hz. The minimum duration for pitch analysis was set to .4 seconds, and the minimum loudness was set to 45 dB. Silence was defined as a period of 1 second or more with sound levels of no more than 60 dB.[4] Sequence Viewer analyzed the mean and standard deviation of pitch from the audio files for the part of each WLT2 interview that corresponded to the first time a question was asked of the respondent for each of the questions examined here. In particular, pitch was measured as the mean fundamental frequency (F_0) over the audio recording for the interviewer's question reading turn. Intonation was measured as the standard deviation of the $\log_{10}$(pitch) over the audio recording for the interviewer's question reading turn, a commonly used measure of intonation (Dijkstra, 2016). Acoustic analyses are available on 7583 question-asking turns for 878 respondents in WLT2. There were no audio recordings at all for three respondents, and no Waveform analyses could be conducted for 21 respondents due to poor quality audio recordings. Three respondents were excluded because the interviewer completed fewer than 10 interviews, yielding an analytic dataset for acoustic measurements of 7560 question readings for 875 respondents.

We standardized the interviewer voice characteristics for male and female interviewers by subtracting the WLT2 gender-specific mean and dividing by the gender-specific standard deviation of each acoustic measurement (Schaeffer et al., 2018). In particular, for question *i* asked by the interviewer of gender *g*, we calculate standardized pitch measures as

$$(MeanPitch_{gi} - \overline{MeanPitch_g})/SD_{MeanPitch_g}$$

and used the same method for standardized intonation. Thus, higher values of pitch and intonation indicate greater deviations from the gender-specific mean in standard deviations units.[5] For female interviewers, the unstandardized mean pitch was 223.6 and intonation was 40.5. For male interviewers, they were 148.8 and 28.0 respectively. Full descriptive statistics are shown in **Table 2.** Analyses of standardized volume measurements are also shown in the online supplement.

Total time on question – WLT2 and CS

Our second dependent variable was the total time spent on each question as measured in paradata. Total time was calculated by subtracting the time that the interviewer entered a particular survey question from the time that they advanced to the next screen for each survey question; the computerized instruments contained only one item per screen. Response times equal to 0 or 1 second reflect likely errors in the paradata and were excluded from the analysis. We excluded all interviewers with fewer than 10 interviews to assist with multilevel model estimation (Van Breukelen & Moerbeek, 2013; Vassallo et al., 2017). This decision excluded one interviewer (3 respondents) in WLT2 and 48 interviewers (218 respondents) in CS. Thus, analytic data sets contain 7970 observations over 899 respondents and 26 interviewers for WLT2 and 9189 observations over 1758 respondents and 92 interviewers for CS. To account for the skewed nature of paradata, we calculated the natural logarithm of the time spent on each question, and trimmed the logged time measure to the first and 99th percentiles (Yan & Olson, 2013). Mean log (response time) was 2.1 for WLT2 and 2.7 for CS.

Paradigmatic sequences – WLT2

Next, we look at respondents to WLT2 with available behavior coded data among interviewers with at least ten interviews (n = 896 respondents; 7905 conversational sequences). Sixteen trained undergraduate coders evaluated each conversational turn on eight different fields. We examined three fields here: the actor (interviewer or respondent), their initial action (e.g., question asking), and an assessment of the initial action (e.g., question read exactly as worded). Two master coders independently coded 10% of the interviews to assess coder reliability.

Table 2. Descriptive statistics, work and leisure today 2 and consumer spending surveys.

	Work and Leisure Today 2		*Consumer Spending*	
	Mean/%	*SD*	*Mean/%*	*SD*
Dependent Variables				
Unstandardized Pitch (F_0)				
Female interviewers	223.6	28.4	n/a	
Male interviewers	148.8	25.1	n/a	
Unstandardized intonation				
Female interviewers	40.5	16.3	n/a	
Male interviewers	28.0	17.8	n/a	
log(Total time spent on each question)	2.1		2.7	
% Paradigmatic sequences	65%		n/a	
Control Variables				
# words in the question	17.3	18.0	26.3	13.1
Interviewer characteristics				
% Female	57%		57%	
% white	54%		68%	
% prior experience	81%		82%	
Within-survey experience	21.2		12.8	
Respondent characteristics				
% Landline	34.1%		42.5%	
Age				
18 to 44	37.8%		43.5%	
45 to 64	36.1%		34.7%	
65+	23.2%		19.9%	
Missing	2.0%		1.9%	
Education				
High school or less	32.5%		39.3%	
Some college	27.0%		20.2%	
BA+	40.0%		39.8%	
Missing	0.5%		0.7%	
Race				
Non-Hispanic white	68.8%		67.2%	
Non-Hispanic black	9.5%		11.5%	
Hispanic	8.5%		13.5%	
Non-Hispanic other	10.5%		6.2%	
Missing	2.8%		1.7%	
Gender				
Male	48.2%		49.4%	
Female	51.8%		50.5%	
Missing	0%		0.14%	

We used these codes to identify whether the exchange between the interviewer and respondent was 'paradigmatic' for each question. We defined a sequence as paradigmatic if the interviewer's initial question was rendered exactly as worded (initial interviewer behavior kappa = 0.93; question asking kappa = 0.64) and followed by a respondent's initial answer that was adequate for the response task (initial respondent behavior kappa = 0.83; answering behavior kappa = 0.79). Additionally, any sequence where the conversation consisted of only three conversational turns, in which the first two were exact reading and adequate answer, followed by neutral feedback from the interviewer (kappa = 0.76), was defined as paradigmatic. Neutral feedback was identified as a short acknowledgment ('thank you'), an affirmation ('ok'), repeating the respondent's answer exactly (R: 'Just two'; I: 'Just the two okay'), laughter only, long motivational feedback ('And again, I really appreciate you helping me out. I think we only need one more for us, for the night, so I do appreciate it'), task related feedback ('Okay. One second. This thing don't wanna work right today. Okay there we go'), time related feedback ('Okay. We're almost there. Just about there.'), or transition statements to the next question ('And um, [clears throat]'). Any interaction that was longer than three conversational turns or was two or three conversational turns that did not meet these definitions was identified as non-paradigmatic. In WLT2, 64.9% of sequences were identified as paradigmatic.[6]

Responses to survey questions – WLT2 and CS

Finally, we examined survey question responses in both WLT2 and CS. These analyses were weighted to account for the dual frame design and nonresponse; standard errors were further adjusted to account for the dual frame stratification and clustering of respondents within interviewers. We looked at differences across versions in (1) item missing data rates and (2) actual responses among respondents who provided a substantive answer.

Control variables

Question characteristics

We controlled for the number of words in the question stem (Mean: WLT2 = 17.3, CS = 26.3). Longer questions will take longer to administer (e.g., Olson & Smyth, 2015), have more opportunities for the interviewer to misread or for the respondent to interrupt (e.g., Olson et al., 2019), and more opportunities for the interviewer to vary their pitch.

Interviewer characteristics

Because the emphasis experiment is constant within interviewers in WLT2, we controlled for interviewer gender (Female: WLT2 and CS = 57%), race (White: WLT2 = 54%, CS = 68%) and within-survey experience, measured by the order in which each interview was completed within each interviewer (i.e., 1 = 1st interview, 2 = 2nd interview, etc.). This count was log-transformed to allow for a non-linear learning effect (e.g., Olson & Peytchev, 2007). We also controlled for overall interviewer experience with the Voxco CATI software in WLT2 (81% 1+ years prior Voxco experience) and general interviewing experience in CS (82% 1+ years general experience).

Respondent characteristics

Cell phone interviews tend to last longer than landline interviews (Timbrook et al., 2018); thus, we controlled for whether the respondent did the interview on a landline (WLT2 = 31.4%, CS = 42.5%) or cell phone. Respondents with lower cognitive abilities, commonly measured by age and education (Knauper, 1999; Krosnick, 1991), may take longer and have a more difficult time answering survey questions. As such, we controlled for age (WLT2/CS: 18 to 44 years = 38%/44%, 45 to 64 years = 36%/35%, 65+ years = 23%/20%, missing = 2%/2%) and education (high school or less = 33%/39%, Some college = 27%/20%, BA+ = 40%/40%, missing = 2%/2%). We also controlled for race (Non-Hispanic White = 69%/ 67%, Non-Hispanic Black = 10%/12%, Hispanic = 9%/14%, Non-Hispanic Other = 11%/6%, Missing = 1%/1%) and gender (Male = 48%/49%, Female = 52%/51%; in WLT2, three cases missing gender were imputed using interviewer assessment of sex) of the respondent to account for any potential differential nonresponse bias over interviewers in these characteristics. We also included the weight

variable as a covariate in both surveys. In WLT2, some of the experimental questions were embedded in skip patterns, and thus we controlled for whether the respondent had ever been employed; owned a car, truck, or other vehicle; or lived in a single person household to account for differential selection into the experimental questions.

Analysis methods

We used cross-classified random effects logistic regression models (Beretvas, 2011; Raudenbush & Bryk, 2002) to simultaneously evaluate the association of question, respondent, and interviewer characteristics with interviewer vocal characteristics, response time, and paradigmatic sequences. Each behavior was cross-classified by respondents and by questions, with questions and respondents nested within interviewers.

Cross-classified models

We predicted the standardized mean pitch and intonation and the log-transformed paradata measure of the number of seconds on each question using cross-classified linear models (Beretvas, 2011). In particular, $Y_{i(j_1 j_2)k}$ is the measured standardized pitch or intonation for observation i representing respondent j_1, question j_2, and interviewer k.

The base model examined the measured pitch or intonation as a function of an overall mean (γ_0) plus random effects due to the respondent ($u_{j_1} \sim N(0, \tau_{uj_1})$), the question ($u_{j_2} \sim N(0, \tau_{uj_2})$), and the interviewer ($\upsilon_k \sim N(0, \tau_{uk})$), and a residual term ($e_{i(j_1 j_2)k} \sim N(0, \sigma^2{}_e)$): $Y_{i(j_1 j_2)k} = \gamma_0 + \upsilon_k) u_{j_1} + u_{j_2} + e_{i(j_1, j_2)k}$. From the base model, we estimated the proportion of the variance in $Y_{i(j_1 j_2)k}$ associated with questions, respondents, and interviewers. For example, we used $\rho_{\text{int}} = \hat{\tau}_{uk} / (\hat{\tau}_{uj_1} + \hat{\tau}_{uj_2} + \hat{\tau}_{uk} + \sigma^2{}_e)$ for the proportion of variance due to interviewers. All of the base models indicated significant interviewer, question, and respondent-related variation in interviewer pitch, intonation, paradigmatic sequences, and response time ($p < .0001$ for all base models).

We then added measures of our experimental condition — the indicator for whether the respondent was assigned to the emphasis condition or not, a categorical measure of the number of words that were emphasized in the item, and an interaction between the number of emphasized

words and the emphasis condition, as well as the number of words in the question, interviewer characteristics, and respondent characteristics:

$$
\begin{aligned}
Y_{i(j_1 j_2)k} = {} & \gamma_0 + \beta_1 EmphasisCond + \beta_2 TwoEmphWords \\
& + \beta_3 ThreeEmphWords + \beta_4 SixEmphWords \\
& + \beta_5 TwoEmphWords * EmphasisCond \\
& + \beta_6 ThreeEmphWords * EmphasisCond \\
& + \beta_7 SixEmphWords * EmphasisCond \\
& + \beta_8 NumWordsStem + \sum_{t=1}^{r} \beta_t Iwer\_char_k \\
& + \sum_{m=1}^{p} \beta_m Respondent\_char_{j_1} \\
& + \upsilon_k + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)k}
\end{aligned}
$$

All of the linear models were estimated using restricted maximum likelihood estimation in Stata 15.0 *mixed* with random intercepts for questions, respondents, and interviewers (Rabe-Hesketh & Skrondal, 2012). For paradigmatic sequences, we modified the model estimation form to a cross-classified random effects logistic regression, modeling the probability that the interaction was paradigmatic(=1) versus not paradigmatic(=0). The modeling framework and strategy were identical to that described above, with the link function modified to a logit link, and the residual variance constrained to $\pi^2/3$. We used Stata 15.1's *meqrlogit* to estimate these cross-classified logistic regression models.

Survey responses

We examined missing data and survey responses, accounting for the unequal selection and nonresponse-adjustment weights and the strata. We used simple bivariate analyses to examine the differences in item nonresponse rates and responses to the survey questions across experimental conditions.

Findings

RQ1: Does emphasizing words in survey questions influence interviewers' vocal characteristics?

We start by estimating base models for mean pitch and intonation over the question asking turns in WLT2, excluding any covariates; these data

are not available for CS. In WLT2, 44.3% of the variation in pitch (τ_{uk} = 0:494, p < :0001) and 21.1% of the variation in intonation (τ_{uk} = 0.221, p < .0001) come from the interviewer, 8.2% of the variation in pitch (τ_{uj2} = 0.092, p < .0001) and 4.9% of variation in intonation (τ_{uj2} = 0.051, p < .0001) is due to the question, and 10.0% of the variation in pitch (τ_{uj1} = 0.111, p < .0001) and 2.7% of the variation in intonation (τ_{uj1} = 0.028, p < .0001) is due to the respondent. The large proportion of variance that is due to the interviewer is unsurprising — these, after all, are interviewer voice characteristics. What is more striking is the substantial variation in pitch and intonation due to the question and respondent, suggesting that interviewers are adapting how they ask questions, at least somewhat, to the question they are asking and to the person to whom they are talking.

We now examine the emphasis experiment in WLT2. Coefficients for the independent variables are shown in **Table 3**. Consistent with H1a, there is no significant difference in interviewer's voice pitch (Models 1 and 2) for questions with and without emphasis (z = −1.05, p = 0.29), or for questions with varying numbers of emphasized words ($\chi^2(3)$ = 5.79, p = 0.12). There is, however, a significant difference in intonation ($\chi^2(3)$ = 52.08, p < .0001 Models 3 and 4). Consistent with H1b, questions with one, three or six emphasized words are asked with significantly *higher levels* of intonation than the same questions without emphasis. Additionally, the effect of emphasis on intonation *is largest* for questions with one emphasized word (marginal effect = 0.27 gender-centered standard deviation units for intonation), compared to questions with more emphasized words. In contrast, counter H1b, questions with two (adjacent) emphasized words are asked with significantly *lower levels* of intonation than the same question without emphasis (marginal effect = −0.10 gender-centered standard deviation units for intonation). The effect of emphasis on intonation is not statistically different between questions with 3 or 6 emphasized words (z = 1.00, p = 0.32; marginal effect 3 words = 0.07 gender-centered standard deviation units for intonation; marginal effect 6 words = 0.14 gender-centered standard deviation units for intonation).

The included covariates explained very little of the interviewer, respondent, or question-level variance in the vocal characteristics. For pitch, including the experimental conditions and control variables *increased* the size of the variance components at the interviewer and question level, and did not change the respondent or residual variance

Table 3. Coefficients and standard errors from cross-classified random effects linear and logistic regression models predicting pitch, intonation, response time, and paradigmatic sequences.

	WLT2 standardized pitch		WLT2 standardized intonation		WLT2 paradigmatic sequence		WLT2 Log(Seconds) on question		CS Log(Seconds) on question	
	Model 1 Coef (SE)	Model 2 Coef (SE)	Model 3 Coef (SE)	Model 4 Coef (SE)	Model 5 Coef (SE)	Model 6 Coef (SE)	Model 7 Coef (SE)	Model 8 Coef (SE)	Model 9 Coef (SE)	Model 10 Coef (SE)
Emphasis version (ref = No emphasis)										
Emphasis	−0.33	−0.28	0.08	0.27	−0.80****	−0.50*	0.08*	0.08*	0.01	−0.05*
	(0.32)	(0.32)	(0.18)	(0.18)	(0.17)	(0.20)	(0.04)	(0.04)	(0.01)	(0.02)
# Emphasized Words (ref = 1 word)										
2 words		−0.06		0.37*		−0.84*		0.22		−0.74****
		(0.30)		(0.19)		(0.29)		(0.16)		(0.15)
3 words		0.11		0.46*		−0.01		0.13		−0.11
		(0.34)		(0.21)		(0.33)		(0.19)		(0.12)
6 words		0.26		0.23		3.27		−1.33		–
		(1.84)		(1.14)		(1.68)		(1.00)		
# Emphasized words *Emphasis condition										
2 words * Emphasis		−0.09*		−0.37****		−0.33*		0.03		0.12**
		(0.04)		(0.05)		(0.15)		(0.02)		(0.04)
3 words * Emphasis		−0.07 –		0.21****		−0.41**		0.00		0.06*
		(0.04)		(0.05)		(0.16)		(0.02)		(0.02)
6 words * Emphasis		−0.04		−0.14		−0.30*		−0.06*		–
	(0.05)	(0.07)	(0.20)	(0.03)						
# words in question	−0.01	−0.01	0.004	0.004	−0.04****	−0.10****	0.02****	0.05**	0.01	0.01
	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.03)	(0.004)	(0.02)	(0.01)	(0.00)
Interviewer characteristics										
Female	−0.09	−0.09	0.05	0.05	0.14	0.14	0.02	0.02	−0.03	−0.02
	(0.36)	(0.36)	(0.20)	(0.20)	(0.20)	(0.20)	(0.05)	(0.05)	(0.03)	(0.03)
Race										
Black	−0.16	−0.16	−0.45*	−0.45*	−0.11	−0.12	0.08	0.08	0.04	0.04
	(0.40)	(0.40)	(0.22)	(0.22)	(0.22)	(0.22)	(0.05)	(0.05)	(0.05)	(0.05)
Hispanic/Latino	0.49	0.49	0.28	0.28	0.29	0.29	−0.01	−0.01	0.00	0.01
	(0.45)	(0.45)	(0.25)	(0.25)	(0.25)	(0.25)	(0.06)	(0.06)	(0.04)	(0.04)
2+ races/Other	−0.20	−0.20	−0.34	−0.34	0.03	0.03	0.19*	0.19*	−0.05	−0.05
	(0.60)	(0.60)	(0.34)	(0.34)	(0.32)	(0.32)	(0.08)	(0.08)	(0.05)	(0.05)
Previous experience 1+ year	0.12	0.12	0.46	0.46	0.19	0.19	−0.12*	−0.12*	0.03	0.03
	(0.45)	(0.45)	(0.25)	(0.25)	(0.24)	(0.24)	(0.06)	(0.06)	(0.04)	(0.04)
Within-survey experience	0.03	0.03	−0.03*	−0.03*	0.03	0.03	−0.06****	−0.06****	−0.03****	−0.03****
	(0.02)	(0.02)	(0.01)	(0.01)	(0.04)	(0.04)	(0.01)	(0.01)	(0.01)	(0.01)

Table 3. Continued.

	WLT2 standardized pitch		WLT2 standardized intonation		WLT2 paradigmatic sequence		WLT2 Log(Seconds) on question		CS Log(Seconds) on question	
	Model 1 Coef (SE)	Model 2 Coef (SE)	Model 3 Coef (SE)	Model 4 Coef (SE)	Model 5 Coef (SE)	Model 6 Coef (SE)	Model 7 Coef (SE)	Model 8 Coef (SE)	Model 9 Coef (SE)	Model 10 Coef (SE)
Respondent characteristics										
Cell phone interview	0.03	0.03	0.05	0.05	0.07	0.07	0.05***	0.05***	0.15****	0.15****
	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)	(0.08)	(0.01)	(0.01)	(0.02)	(0.02)
Age										
45–65	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.00	0.00
	(0.04)	(0.04)	(0.03)	(0.03)	(0.09)	(0.09)	(0.01)	(0.01)	(0.01)	(0.01)
65+	0.10*	0.10*	0.00	0.00	−0.34***	−0.34***	0.08****	0.08****	0.08****	0.08****
	(0.04)	(0.04)	(0.04)	(0.04)	(0.10)	(0.10)	(0.02)	(0.02)	(0.02)	(0.02)
Missing	−0.01	−0.01	−0.06	−0.06	−0.04	−0.04	0.002	0.002	0.08*	0.08*
	(0.08)	(0.08)	(0.07)	(0.07)	(0.20)	(0.20)	(0.03)	(0.03)	(0.03)	(0.03)
Respondent Sex										
Female 0.07*	0.07*	0.09****	0.09****	0.12	0.12	−0.03	−0.03	0.00	0.00	
	(0.03)	(0.03)	(0.02)	(0.02)	(0.07)	(0.07)	(0.01)	(0.01)	(0.01)	(0.01)
Missing	–	–	–	–	–	–	–	–	0.15	0.15
									(0.10)	(0.10)
Education										
Some college	0.06	0.06	−0.01	−0.01	−0.06	−0.06	0.01	0.01	−0.03	−0.03
	(0.04)	(0.04)	(0.03)	(0.03)	(0.09)	(0.09)	(0.01)	(0.01)	(0.02)	(0.02)
BA+	0.03	0.03	0.01	0.01	0.08	0.08	−0.03*	−0.03*	−0.05****	−0.05****
	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)	(0.08)	(0.01)	(0.01)	(0.01)	(0.01)
missing	−0.15	−0.15	−0.07	−0.07	−0.10	−0.10	0.07	0.07	−0.20****	−0.20****
	(0.22)	(0.22)	(0.19)	(0.19)	(0.48)	(0.49)	(0.08)	(0.08)	(0.05)	(0.05)
Race										
Non-Hispanic Black	0.04	0.04	0.00	0.00	−0.11	−0.11	0.04	0.04	0.09****	0.09****
	(0.05)	(0.05)	(0.04)	(0.04)	(0.12)	(0.12)	(0.02)	(0.02)	(0.02)	(0.02)
Hispanic	0.07	0.08	0.03	0.03	−0.20	−0.20	0.07**	0.07**	0.07**	0.07**
	(0.06)	(0.06)	(0.05)	(0.05)	(0.14)	(0.14)	(0.02)	(0.02)	(0.02)	(0.02)
Non-Hispanic Other	0.11*	0.11*	0.04	0.04	−0.24*	−0.24*	0.07**	0.07**	0.03	0.03
	(0.05)	(0.05)	(0.04)	(0.04)	(0.12)	(0.12)	(0.02)	(0.02)	(0.02)	(0.02)
Missing	0.03	0.03	0.10	0.10	−0.58*	−0.59*	0.07	0.07	0.01	0.01
	(0.10)	(0.10)	(0.08)	(0.08)	(0.23)	(0.23)	(0.03)	(0.04)	(0.03)	(0.03)
Ever employed	0.01	0.01	0.05	0.05	−0.08	−0.08	−0.01	−0.01	–	–
	(0.08)	(0.08)	(0.07)	(0.07)	(0.19)	(0.19)	(0.03)	(0.03)		
Own car/truck	−0.02	−0.02	0.01	0.01	0.08	0.08	−0.07****	−0.07****	–	–
	(0.05)	(0.05)	(0.04)	(0.04)	(0.11)	(0.11)	(0.02)	(0.02)		

Table 3. Continued.

	WLT2 standardized pitch		WLT2 standardized intonation		WLT2 paradigmatic sequence		WLT2 Log(Seconds) on question		CS Log(Seconds) on question	
	Model 1 Coef (SE)	Model 2 Coef (SE)	Model 3 Coef (SE)	Model 4 Coef (SE)	Model 5 Coef (SE)	Model 6 Coef (SE)	Model 7 Coef (SE)	Model 8 Coef (SE)	Model 9 Coef (SE)	Model 10 Coef (SE)
Single person HH	0.00	0.00	0.00	0.00	−0.13	−0.13	0.02	0.02	–	–
	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)	(0.08)	(0.01)	(0.01)		
Base weight	0.02	0.02	−0.01	−0.01	−0.03	−0.03	−0.005	−0.005	−0.01	−0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.07)	(0.07)	(0.01)	(0.01)	(0.01)	(0.01)
Intercept	0.17	0.15	−0.22	−0.22	1.82****	2.75****	1.76****	1.36****	2.31****	2.70****
	(0.45)	(0.45)	(0.21)	(0.21)	(0.40)	(0.43)	(0.12)	(0.19)	(0.28)	(0.14)
Variance components										
Interviewer	0.58****	0.58****	0.18****	0.18****	0.14****	0.14****	0.01****	0.01****	0.01****	0.01****
Question	0.12****	0.12****	0.05****	0.04****	0.37****	0.09****	0.06****	0.04****	0.08****	0.01****
Respondent	0.11****	0.11****	0.03****	0.03****	0.29****	0.29****	0.01****	0.01****	0.01****	0.01****
Residual	0.42	0.42	0.75	0.74	–	–	0.11	0.11	0.15	0.15
Likelihood ratio test for variance components	4957.86****	4959.32****	1332.72****	1181.40****	582.21****	232.24****	3063.81****	1530.26****	3033.31****	798.36****
Log-likelihood	−8053.91	−8057.94	−9846.29	−9825.21	−4366.27	−4356.43	−2901.07	−2903.03	−4843.42	−4841.33
Model Wald $\chi 2$	31.25	37.02	35.26	91.56****	77.17****	142.94****	235.68****	271.23****	237.57****	280.84****
AIC	16171.82	16185.89	19750.59	19270.43	8788.54	8780.86	5860.14	5876.07	9740.85	9744.65
# observations	7560	7560	7560	7560	7905	7905	7970	7970	9189	9189
# respondents	875	875	875	875	896	896	899	899	1758	1758

Interviewers with at least 10 interviews.

* p < .05 ; ** p < .01 ; *** p < .001 ; **** p < .0001

Models 1 through 4 and 7 through 10 are linear regression models. Models 5 and 6 are logistic regression models.

components. For intonation, the included covariates explained just under 20% of the interviewer-level variance, 12% of the question-level variance, and none of the respondent-level or residual variance.

RQ2: Does emphasizing words affect interviewer/respondent behaviors (Paradigmatic sequences and response time) during the interview?

For paradigmatic sequences in WLT2, 18% of the variance is at the question level, followed by 7% at the respondent level and 6.1% at the interviewer level. For response time, in both WLT2 and CS, the largest proportion of total variance is due to question-related factors (63.0% in WLT2; 33.7% in CS) with respondents accounting for about 5% and interviewers between about four and six percent.

We now examine the association between emphasis and whether or not the interviewer/ respondent interaction was paradigmatic in WLT2. Across all questions, conversational sequences on items with emphasis were less likely to be paradigmatic than on items without emphasis, consistent with H2b (Table 3, Models 5 and 6). Examining the marginal effects of emphasis reveals that asking and answering questions paradigmatically for questions without emphasis occurs 76.1% of the time, compared to 59.0% of the time for questions with emphasis. This holds across the number of words that are emphasized ($\chi^2(3) = 7.16$, $p = 0.067$).

Turning to response time, the effect of the emphasis experiment in WLT2 (Table 3, Models 7 and 8) is clear – questions with emphasis take longer than questions without emphasis, consistent with H2d. This occurs overall (coef = 0.084, p = 0.035), and is modified by the number of words that are emphasized ($\chi^2(3) = 9.87$, $p = 0.020$). The marginal effect of emphasis is 0.08 log-seconds ($e^{0.084} = 1.09$ seconds) longer, for questions with one emphasized word; 0.11 log-seconds (1.11 seconds) longer for those with two emphasized words; 0.08 log-seconds (1.08 seconds) longer for questions with three emphasized words; and 0.03 log-seconds (1.02 seconds) longer for questions with 6 emphasized words. In CS (Table 3, Models 9 and 10), the effect is more complicated. Questions with one emphasized word (both income questions, thus confounding number of words with content) take *less time* (about 0.95 seconds less) than the same questions without emphasis, consistent with H2c (coef

= −0.048, p = 0.032). Questions with two emphasized words take *longer* than the same questions with no emphasis (coef = 0.118, p = 0.001; marginal effect = 0.07 log seconds, or 1.07 seconds), consistent with H2d and, reassuringly, consistent in both direction and magnitude with the questions with two emphasized words in WLT2 (coef = 0.118, p = 0.001). Additionally consistent with H2d, questions with three emphasized words *take slightly longer* than the same questions without words emphasized (coef = 0.055, p = 0.024; marginal effect = 0.01; 1.01 seconds longer). Thus, across both WLT2 and CS, questions with emphasis generally take longer than those without.

Across the paradigmatic sequences and response timing outcomes, the included covariates explained over 85% of the question-level variance from the base model. This is notable because there were very few question-level covariates included. Between about one-quarter and one-half of the interviewer-related variance in these outcomes was explained by the included covariates, and between about 10% and 30% of the respondent-level variance was explained.

RQ3: Does emphasizing words affect respondents' answers?

We start with item nonresponse rates. In WLT2, item nonresponse rates are too low (less than 1% in most items) to detect a difference across experimental conditions for the individual items. Likewise, contrary to both H3a and H3b, we found no significant difference across the emphasis conditions in whether the respondent failed to answer *any* of the questions included as part of this experiment (emphasis = 3.85%, no emphasis = 5.02%, F(1,50) = 0.57, p = 0.45). In CS, in contrast, consistent with H3a, two of the six items (Q1 and Q2) had reduced item nonresponse rates in the emphasis condition ($p < 0.05$), but consistent with H3b, one item (D9) had an increased item nonresponse rate in the emphasis condition ($p < 0.05$). Thus, in general, emphasis either has no effect on item nonresponse rates or slightly reduces them.

Finally, we examine whether emphasis affected respondent's reports to the survey questions themselves (**Table 4**). We find little evidence for such effects. Of the 16 items, responses differed significantly across emphasis conditions in only one (WLT2 Q32, 5.1% had a parking ticket in the emphasis condition compared to 10.8% in the no emphasis condition). With so many tests, it is difficult to attribute this one significant difference to anything other than Type I error.

Table 4. Survey responses by emphasis condition, WLT2 and CS.

Survey	*Question stem*	*No emphasis*	*Emphasis*	*Design-based F*
TIME FRAME				
WLT2	Q2. Compared to 10 YEARS AGO IN 2005, do you think people have more leisure time, less leisure time or about the same amount?			
	More	16.11%	13.02%	1.21
	Same	33.36%	29.31%	
	Less	50.53%	57.67%	
WLT2	Q5. Have you EVER been employed for pay or profit?			
	Yes	95.32%	88.81%	3.34
	No	4.68%	11.19%	
WLT2	Q6. Have you EVER been laid off from a job?			
	Yes	34.93%	38.60%	0.92
	No	65.07%	61.40%	
WLT2	Q7. Have you EVER been fired from a job?			
	Yes	17.12%	18.37%	0.09
	No	82.88%	81.63%	
CS	Q11. Have you EVER participated in a sharing economy service as either an OWNER or RENTER?			
	Yes	23.89%	26.43%	0.71
	No	76.11%	73.57%	
WLT2	Q32. During THE LAST YEAR, how many parking tickets have you received?			
	Zero	89.25%	94.94%	4.66*
	1 or more	10.75%	5.06%	
WLT2	Q33. During THE LAST YEAR, how many speeding tickets have you received?			
	Zero	94.82%	90.30%	3.55
	1 or more	5.18%	9.70%	
CS	D45. What is your total MONTHLY household income, before taxes?			
	Mean	7.14	7.41	1.35
CS	D46. Is your total MONTHLY household income before taxes $4000 or more, or is it less than $4000?			
	Mean	6.97	6.37	2.94
REFERENT OF QUESTION				
WLT2	Q24. The next questions are about mobile phones. For these questions, a cell phone is a mobile telephone on which only calls or texts are made and received. A smartphone is a mobile telephone on which the user can access the internet, use apps, and read email, as well as send and receive calls and texts. Do YOU happen to have a working CELL PHONE OR A SMARTPHONE?			
	Cell phone	23.73%	26.56%	0.21
	Smartphone	66.15%	63.68%	
	Both	2.79%	2.57%	
	None	7.34%	7.20%	
WLT2	Q25. Does ANYONE ELSE in your household have a working mobile phone, either cell phone or smartphone?			
	Yes	69.25%	72.37%	0.80
	No	30.75%	27.63%	
INCLUSION/EXCLUSION				
WLT2	Q42. How many people, INCLUDING YOURSELF, live in your household?			
	Mean	2.58	2.67	0.91
WLT2	Q43. How many people, INCLUDING YOURSELF, are adults age 18 and older?			
	Mean	2.41	2.34	−0.81
CS	D9. INCLUDING YOURSELF, how many adults, 18 years of age or older, live in this household?			
	Mean	2.28	2.22	−0.60

Continued

Table 4. Continued

Survey	*Question stem*	*No emphasis*	*Emphasis*	*Design-based F*
RESPONSE OPTIONS				
CS	Q1. Please think about your total household spending over the past four weeks, and all the purchases you have made. Compared to the same time a year ago, would you say you are spending MORE money, spending LESS money, or about the SAME amount of money?			
	More	36.21%	36.44%	0.63
	Less	17.57%	20.13%	
	Same	46.22%	43.43%	
CS	Q2. Compared to a year ago, do you feel MORE willing to spend money on things you may not need, LESS willing to spend money, or do you feel the SAME about spending money on things you may not need?			
	More	11.49%	10.03%	0.26
	Less	49.43%	50.31%	
	Same	39.08%	39.66%	

Discussion

Despite its ubiquity in survey questionnaires, adding emphasis to survey questions has very little effect on survey measurement quality in this study. **Table 5** summarizes the hypotheses and findings across all of the analyses from these emphasis experiments. Interviewers implement emphasis by changing the intonation of their voice as they ask survey questions, but how they do this depends on the number of words emphasized. It appears to be easier to vocally emphasize one word than two or more. This emphasis changes the interaction between interviewers and respondents by leading to longer and less paradigmatic interactions, and again appears to depend on the number of words that are emphasized in the question. So, interviewers inconsistently administer questions with emphasis and the use of emphasis appears to change the interaction between interviewers and respondents, but not for the better. The use of emphasis has modest to no effect on item nonresponse rates in these surveys, and has no effect on answers to the survey questions asked here. With this, there is little here to suggest that questionnaire designers should use all caps emphasis when writing survey questions for interviewer-administered telephone surveys.

We are able to explain a good amount of variation in question response time and presence of paradigmatic sequences with the included

Table 5. Summary of hypotheses and findings.

Outcome	*Hypothesis*	*Result*
Interviewer voice		
Pitch	H1a: Similar mean pitch with emphasis	No effect
Intonation	H1b More intonation with emphasis	Yes, depending on # emphasized words
Interviewer-Respondent Interactions		
Paradigmatic interactions	H2a: Emphasis more likely to be paradigmatic	Emphasis less likely to be paradigmatic
	H2b: Emphasis less likely to be paradigmatic	
Response timing	H2c: Emphasis has shorter response times	Emphasis has longer response times, depending slightly on # emphasized words
	H2d: Emphasis has longer response times	
Item Nonresponse	H3a: Emphasis will have lower item nonresponse rates	No effect or less item nonresponse
	H3b: Emphasis will have higher item nonresponse rates	
Survey Responses		
Time frame = 10 years ago	H3c1: No effect	No effect
Time frame = Ever	H3c2: Higher reports in emphasis	No effect
Time frame = Last year	H3c3: Lower reports in emphasis	One question lower reports in emphasis
Time frame = Monthly	H3c4: Lower reports in emphasis	No effect
Referent of question = You	H3d1: Lower reports in emphasis	No effect
Referent of question = Anyone else	H3d2: Higher reports in emphasis	No effect
Inclusion = Including yourself	H3e: Higher reports in emphasis	No effect
Response options	H3f: No effect	No effect

covariates. But, significant variance across interviewers, questions, and respondents remains, especially in interviewer pitch and intonation. It is interesting that there is significant variation in these interviewer vocal characteristics across question administrations and across respondents — interviewers appear to adapt their voice depending on *what* they are saying and *who* they are talking to. Future work should explore this in more detail.

Additional analyses show that the interviewers in the emphasis condition are notably more likely to verify the respondent's answer than the interviewers in the non-emphasis condition, leading to longer and non-paradigmatic interactions. They are not more likely to clarify the question or to use probes. For instance, in response to the question about the number of speeding tickets received during the last year, a respondent reported "None," which was then verified by the interviewer saying "Zero." Thus, it appears that the use of emphasis cues interviewers to double check the respondent's answer. These additional conversational turns lengthen the interaction, but do not lead to differences in

the distribution of answers to the survey questions. We note that interviewers were *not* specifically trained to do this extra verification in the emphasis condition; something about the emphasis communicated to the interviewers the need for more attention, at least in WLT2.

This study has limitations. Although the emphasis experiment was replicated across two survey organizations, the questionnaire differed between these two organizations, as did the experimental design (between-respondents for WLT2 and within-respondents for CS). It is reassuring that many of the results replicated across these two studies. However, some differences between the two surveys remained, such as differences in item nonresponse rates for CS but not WLT2 and differences in the effect of the number of emphasized words on response timing (only income questions had one emphasized word in CS). Unfortunately, given the differences in design, we cannot disentangle how much of these differences are due to the questions themselves versus the survey organizations versus the experimental design. Additionally, we do not have acoustic analyses and behavior codes for CS; thus, we cannot replicate these analyses. Future research should explore how a within-interviewer/respondent or between-interviewer/respondent design affects conclusions about emphasis in interviewer-administered surveys using identical questionnaires. Future research should also specifically vary the number of emphasized words on the same question to disentangle content from how many words are emphasized in a question stem. The number of interviewers at each organization involved in these experiments was small. As such, replication with a larger interviewer corps would benefit our understanding of this phenomenon more generally. Additionally, at both of these organizations, standard practice was to use all caps for emphasis, as was done in these surveys. We have no reason to expect that interviewers would use different vocal inflection with underlining, bolding, or italics emphasis, but future work could empirically evaluate this question, including whether there is variation in vocal inflection across different types of emphasis within the same survey. Future work could also examine how emphasis is implemented and whether it is related to survey responses in questionnaires in languages other than English.

We also do not evaluate how emphasis in question stems affects responses to self-administered questionnaires. One might speculate that the effect of emphasis to survey responses and response timing should

be larger in self-administered questionnaires than in interviewer-administered questionnaires, given that there is no mediator who needs to communicate the emphasized words to the respondent. Future research should examine emphasis in question stems for self-administered surveys in more detail. Other types of questions may be more sensitive to an emphasis effect. For example, none of our questions contained an experimental manipulation of emphasis on the word 'not' – this is a fertile area for future research.

We also do not have measures that allow us to evaluate other measures of data quality, including accuracy of responses. One goal of adding emphasis on behavioral questions is to help respondents understand the reference period (e.g., EVER), which may result in more accurate reports. As there were no differences in the reports themselves across the emphasis conditions examined here, we have no reason to expect there would be differences in accuracy rates. Yet future research is needed to empirically evaluate this issue.

Adding time frame, referent, and response option emphasis to the questions examined here in interviewer-administered surveys had little effect on survey data quality and lengthened the interaction between interviewers and respondents. The results of this study run counter to the common questionnaire design recommendation that emphasis on important words in survey questions helps respondents. As the first study of its kind, more research is needed to examine the role that these and other types of emphasis plays in a variety of other questions.

Notes

1. Other forms of emphasis may highlight a difference from a preceding question.
2. Longer time frames also pose more difficult recall tasks, and more salient behaviors are easier to recall (Tourangeau et al., 2000). In this study, the length of recall was not experimentally manipulated, as our focus is on how emphasis affects reports within questions with a variety of time frames.
3. The no emphasis condition had 225 landline respondents and 226 cell phone respondents. The emphasis condition had 226 landline respondents and 225 cell phone respondents.
4. Patel and Broughton (2002) estimate that call center employees are exposed to phone calls that range from 65–88 db for a particular headset; as such 60 dB is lower than this range. Jefferson (1989) identifies one second of silence as the conversational threshold for tolerance of silence; as such, 1 second or more exceeds this value. The values used here also are the Sequence Viewer defaults.
5. Although other examinations of interviewer acoustic measurements standardize the interviewers voice characteristics of individual sounds relative to other words in a single conversational

turn (e.g., Benki et al., 2011), we have multiple questions and substantial interviewer variability in how each of these questions are asked. Moreover, as one of the items that is part of the emphasis experiment occurs as the first question of the survey, no questions occur prior to the respondent hearing the item with emphasis and thus there is no logical 'control' question.

6. As a robustness check, we also relaxed our requirement for exact question reading, conducting all of the analyses including question readings read with stutters but no changes to the question wording as paradigmatic, raising the percent of paradigmatic sequences to 68.4%. Because here are no meaningful changes in the model results, we keep the stricter definition in our analyses.

Acknowledgments An earlier version of this paper was presented at the Midwest Association for Public Opinion Research annual meeting, November 2016, Chicago, IL, and at the Groningen Symposium on Language and Interaction, January 2017, University of Groningen. We thank Amanda Ganshert for research assistance.

Funding This material is based upon work supported by the National Science Foundation under [Grant No. SES-1132015]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Notes on contributors

Kristen Olson is the Leland J. and Dorothy H. Olson Professor in the Department of Sociology and the Director of the Bureau of Sociological Research at the University of Nebraska-Lincoln. Her research focuses on nonsampling errors in household surveys. Her current projects examine interviewer effects in telephone surveys, within-household selection in mail surveys, and questionnaire design issues in mail, web, and mobile surveys. She has published work on these topics in sociological, methodological, and public opinion research journals.

Jolene D. Smyth is a Professor and Chair of the Department of Sociology at the University of Nebraska-Lincoln. Her research focuses on improving data collection processes through the reduction of measurement and nonresponse error. Her current projects examine how questionnaire design impacts interviewer/respondent interactions and responses in telephone surveys, how various questionnaire design features (e.g., question wording, visual emphasis, response scales, etc.) impact responses in both interviewer and self-administered surveys, and mobile web survey design.

References

Benki, J., Broome, J., Conrad, F., Groves, R., & Kreuter, F. (2011). *Effects of speech rate, pitch, and pausing on survey participation decisions* [Paper presentation]. The American Association for Public Opinion Research Annual Meeting, Phoenix, AZ.

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). Routledge.

Birch, S., & Clifton, C. (2002). Effects of varying focus and accenting of adjuncts on the comprehension of utterances. *Journal of Memory and Language*, *47*(4), 571–588. https://doi.org/10.1016/S0749-596X(02)00018-9

Bureau of Labor Statistics (BLS). (2014). *How the government measures unemployment* (Technical Documentation). U.S. Bureau of Labor Statistics. Current Population Survey (CPS). https://www.bls.gov/cps/cps_htgm.pdf

Carlson, K., Frazier, L., & Clifton, C., Jr.. (2009). How prosody constrains comprehension: A limited effect of prosodic packaging. *Lingua*, *119*(7), 1066–1082. https://doi.org/10.1016/j.lingua.2008.11.003

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142–2156. https://doi.org/10.1121/1.392372

Crawford, S., McCabe, S. E., & Pope, D. (2005). Applying web-based survey design standards. *Journal of Prevention & Intervention in the Community*, *29*(1–2), 43–66. https://doi.org/10.1300/J005v29n01_04

Dijkstra, W. (2016). *Sequence viewer version 6.1*. http://www.sequenceviewer.nl/downloads/Reference%20SV6.1.pdf

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons, Inc.

Doyle, J. R., & Bottomley, P. A. (2009). The massage in the medium: Transfer of connotative meaning from typeface to names and products. *Applied Cognitive Psychology*, *23*(3), 396–409. https://doi.org/10.1002/acp.1468

Duncan, S., & Rosenthal, R. (1968). Vocal emphasis in experimenters' instruction reading as unintended determinant of subjects' responses. *Language and Speech*, *11*(1), 20–26. https://doi.org/10.1177/002383096801100103

Falcone, B. (2017, May 18–21). *Does typographic cuing improve the processing of information from survey questions on a mobile device* [Paper presentation]. The American Association for Public Opinion Research annual conference, New Orleans, LA.

Falcone, B., Malakhoff, L., & Wang, L. (2018, May 16–19). *Italicizing optional instructions on mobile online surveys improves visual filtering of survey content: An eye tracking study* [Paper presentation]. The American Association for Public Opinion Research annual conference, Denver, CO.

Fraundorf, S. (2012). *What happened (and what didn't): Prominence promotes representation of salient alternatives in discourse* [Ph.D.]. University of Illinois at Urbana-Champaign.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how contrastive contrastive accenting really is. *Journal of Memory and Language*, *63*(3), 367–386. https://doi.org/10.1016/j.jml.2010.06.004

Ito, K., & Speer, S. R. (2006). *Immediate effects of intonational prominence in a visual search task* [Paper presentation]. The Speech Prosody 2006, Third International Conference, Dresden, Germany.

Jefferson, G. (1989). Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary perspective*. Multilingual Matters. (pp. 166–196). (Expanded version in Tilburg Papers in Language and Literature, No. 42, 1–83 (1983)).

Knauper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*, *63*(3), 347–370. https://doi.org/10.1086/297724

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

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, *32*(4), 451–454. https://doi.org/10.1121/1.1908095

Margolin, S. J. (2013). Can bold typeface improve readers' comprehension and metacomprehension of negation?. *Reading Psychology*, *34*(1), 85–99. https://doi.org/10.1080/02702711.2011.626107

McAteer, E. (1992). Typeface emphasis and information focus in written language. *Applied Cognitive Psychology*, *6* (4), 345–359. https://doi.org/10.1002/acp.2350060406

McCarthy, M., & Mothersbaugh, D. (2002). Effects of typographic factors in advertising-based persuasion: A general model and initial empirical tests. *Psychology & Marketing*, *19*(7–8), 663–691. https://doi.org/10.1002/mar.10030

National Center for Education Statistics (NCES). (2012). Early childhood longitudinal study: Kindergarten class of 2010–2011. *Parent Interview*. https://nces.ed.gov/ecls/pdf/kindergarten2011/Fall_K_Parent_Interview.pdf

National Center for Science and Engineering Statistics (NCSES). (2016). *Survey of earned doctorates questionnaire*. National Science Foundation. https://www.nsf.gov/statistics/srvydoctorates/surveys/srvydoctorates_2016.pdf

Olson, K., & Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly*, *71*(2), 273–286. https://doi.org/10.1093/poq/nfm007

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

Patel, J. A., & Broughton, K. (2002). Assessment of the noise exposure of call centre operators. *Annals of Occupational Hygiene*, *46*(8), 653–661. https://doi.org/10.1093/annhyg/mef091

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata, third edition, volume II: Categorical responses, counts, and survival* (3rd ed. ed.). Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Redline, C. (2011). *Clarifying survey questions* [Unpublished Doctoral Dissertation. Joint Program in Survey Methodology]. University of Maryland.

Redline, C. (2013). Clarifying categorical concepts in a web survey. *Public Opinion Quarterly*, *77*(S1), 89–105. https://doi.org/10.1093/poq/nfs067

Ryalls, J., & Behrens, S. (2000). *Introduction to speech science: From basic theories to clinical applications*. Allyn and Bacon.

Sanford, A. J. S., Sanford, A. J., Molle, J., & Emmott, C. (2006). Shallow processing and attention capture in written and spoken discourse. *Discourse Processes*, *42*(2), 109–130. https://doi.org/10.1207/s15326950dp4202_2

Schaeffer, N. C., & Maynard, D. W. (1996). From paradigm to prototype and back again: Interactive aspects of cognitive processing in standardized survey interviews. In N. Schwarz & S. Seymour (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 65–88). Jossey-Bass /Pfeiffer.

Schaeffer, N. C., Min, B. H., Purnell, T., Garbarski, D., & Dykema, J. (2018). Greeting and response: Predicting participation from the call opening. *Journal of Survey Statistics and Methodology*, *6*(1), 122–148. https://doi.org/10.1093/jssam/smx014

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. Jossey-Bass Publishers.

The American Association for Public Opinion Research (AAPOR). (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.).

Timbrook, J., Smyth, J., & Olson, K. (2018). Why do mobile interviews last longer? A behavior coding perspective. *Public Opinion Quarterly*, *82*(3), 553–582. https://doi.org/10.1093/poq/nfy022

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.

Van Breukelen, G., & Moerbeek, M. (2013). Design considerations in multilevel studies. In M. A. Scott, J. S. Simonoff, & B. D. Marx Eds., *The SAGE handbook of multilevel modeling* (Ch. 11, pp. 183–200). SAGE Publications.

Vassallo, R., Durrant, G., & Smith, P. (2017). Separating interviewer and area effects by using a cross-classified multilevel logistic model: Simulation findings and implications for survey designs. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *180*(2), 531–550. https://doi.org/10.1111/rssa.12206

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–96). John Wiley & Sons.