

# **The effect of question characteristics, respondents and interviewers on response timing and question misreading in CATI surveys**

Kristen Olson  
Jolene Smyth  
Antje Kirchner

University of Nebraska-Lincoln  
NCRN – Virtual Seminar, March 2<sup>nd</sup>, 2016

# Acknowledgments

- This work was supported by the National Science Foundation 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.
- Thanks to Beth Cochran, Amanda Ganshert, Jinyoung Lee, and Jerry Timbrook for research assistance!

# Background

- Questionnaire designers have to make decisions about
  - Question content and wording
  - Question length
  - Ordering of questions
  - Grouping of questions in a battery
  - Offering definitions
  - Instructing interviewers to probe
  - Etc.
- Empirical evaluations of the effects of these design features look at one feature at a time

# Background (2)

- However, respondents and interviewers experience multiple question features simultaneously
  - The joint effects of these question features should be evaluated in tandem.
- Existing work that has looked at multiple questionnaire features simultaneously
  - Web: Yan and Tourangeau (2008)
  - CAPI/ACASI: Couper and Kreuter (2013)
- Joint features of CATI questionnaires have received limited attention

# Background (3)

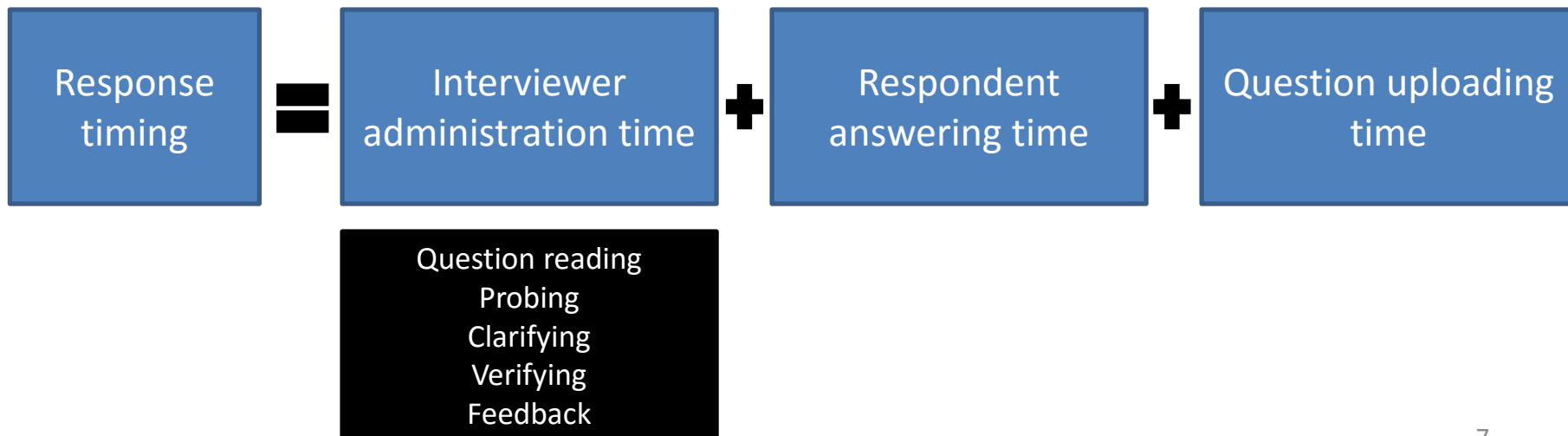
- Additionally, visual design of CATI questionnaires has received limited empirical attention experimentally or otherwise.
- Open question: How do questionnaire features – both structural and visual – affect data quality in CATI surveys?

# Measuring Data Quality

- **Response latencies** (Olson and Parkhurst 2013; Yan and Olson 2013)
  - Common metric that can be used across all survey questions (unlike measures such as acquiescence, straightlining)
  - Has enough variability in interviewer-administered surveys to be evaluated (unlike item nonresponse)
  - Can be easily measured via **paradata** in computer-administered surveys
- **Question misreading** (e.g., Dykema, Lepkowski and Blixt 1997; Mangione, Fowler and Louis 1992; Ongena and Dijkstra 2006)
  - Question is read exactly as written or with changes by adding or omitting words
  - Can be used across all survey questions
  - Can be measured via **behavior codes** in computer-administered surveys

# Response Latencies

- Answering too fast or too slowly may indicate problems in answering, and thus potential data quality issues (Couper 1998; Yan and Olson 2013; Malhotra 2008; Bassili & Fletcher 1991, Yan & Tourangeau 2008, Couper & Kreuter 2013)
  - See Olson and Smyth (2015), JSSAM
- In CATI surveys



# Question Reading Behaviors

- Reading questions exactly as written is important because it decreases the potential for (interviewer-related) measurement error.
  - Although empirical evidence on this is inconclusive (Schaeffer and Dykema, 2011)
- Deviations in question-asking may trigger interactional problems in later stages of the survey (Holbrook, et al. 2015).
- Strong association between reading questions exactly as worded and other interviewer traits such as probing or recording responses (Fowler and Mangione 1990).



# What might affect response latencies and question reading behaviors?

- Question characteristics
  - Necessary question characteristics
  - Respondent task complexity
  - Processing efficiency
- Visual design features
  - Interviewer task complexity

# Necessary question characteristics

	Response Latencies	Exact reading	Major Misreading
<b>Ordering of questions in questionnaire</b> (e.g., Couper and Kreuter 2013; Yan and Tourangeau 2008; but Holbrook et al. 2015)	↓	↓	↑
<b>Length of question</b> (e.g., Yan & Tourangeau 2008; Childs and Landreth 2006; Ongena and Dijkstra 2007)	↑	↓	↑
<b># of response options</b> (e.g, Yan & Tourangeau 2008, but Krosnick & Fabrigar 1997; Fowler and Mangione 1990, Childs and Landreth 2006; but Holbrook et al. 2015)	↑	?	?

# Necessary question characteristics

	Response Latencies	Exact Reading	Major Misreading
<p>Type of question (e.g., Bassili &amp; Fletcher 1991; Tourangeau, Rips, &amp; Rasinski. 2000; Yan &amp; Tourangeau 2008)</p> <ul style="list-style-type: none"> <li>• Attitude</li> <li>• Behavior</li> <li>• Demographics</li> </ul>	<p>↓</p> <p>↑</p> <p>↓</p>	<p>?</p> <p>?</p> <p>↑ ↓</p>	<p>?</p> <p>?</p> <p>↑ ↓</p>
<p>Format of response options (e.g., Smyth, et al. 2009; Couper, et al. 2011; Couper and Kreuter 2013; Childs and Landreth 2006; Mathiowetz and Cannell 1980)</p> <ul style="list-style-type: none"> <li>• Open-ended (numeric vs. text)</li> <li>• Closed-ended nominal</li> <li>• Closed-ended ordinal</li> <li>• Yes-no</li> </ul>	<p>↓</p> <p>↑</p> <p>↓</p> <p>↓</p>	<p>--</p> <p>--</p> <p>↑</p> <p>↑</p>	<p>--</p> <p>--</p> <p>?</p> <p>?</p>

# Respondent task complexity

	Response Latencies	Exact reading	Major Misreading
Question reading level (e.g., Yan & Tourangeau 2008; Fowler and Mangione 1990; Ongena and Dijkstra 2007)	↑	↓	↑
Mismatch between question and response options (e.g., Dykema, Schaeffer, Garbarski, 2016)	↑	↓	↑
Unknown terms (Fowler, 1992)	↑	?	?
Sensitive questions (e.g. Ongena and Dijkstra 2007; but Fowler and Mangione 1990; Holbrook et al. 2015)	↓	?	?

# Processing efficiency

	Response Latencies	Exact reading	Major Misreading
First question in battery (Dykema, Schaeffer and Garbarski, 2016)	↑	↓	↑
Later question in battery (Dykema, Schaeffer and Garbarski, 2016)	↓	↓	↑
Questions with definitions (Fowler, 1992)	↑	↓	↑
Transition statement (Edwards, Schneider, and Brick, 2008)	↓	↓	↑
Feeder question	--	?	?
Follow-up question in a skip	↓	--	--

# Visual Design of Screen

	<b>Response Latencies</b>	<b>Exact reading</b>	<b>Major Misreading</b>
Interviewer instructions (but see Couper and Kreuter 2013)	↑	?	?
Probe on screen	↑	↓	↑
Use of parentheses in question stem (Dykema, Schaeffer, Garbarski, 2016)	↑	↓	↑
Questions asked over two screens (Edwards, Schneider, and Brick, 2008)	↑	--	--
Questions with visual emphasis	↑	--	--

# Respondent and interviewer characteristics

## Respondent

- Age, education as proxies for cognitive abilities (Krosnick 1991; Narayan & Krosnick 1996; Knauper 1999)
- Include sex to account for differential sex composition in telephone surveys
- Computer status and employment status proxy for questionnaire burden

## Interviewer

- Experience, workload, and cooperation rate
- Include sex and race

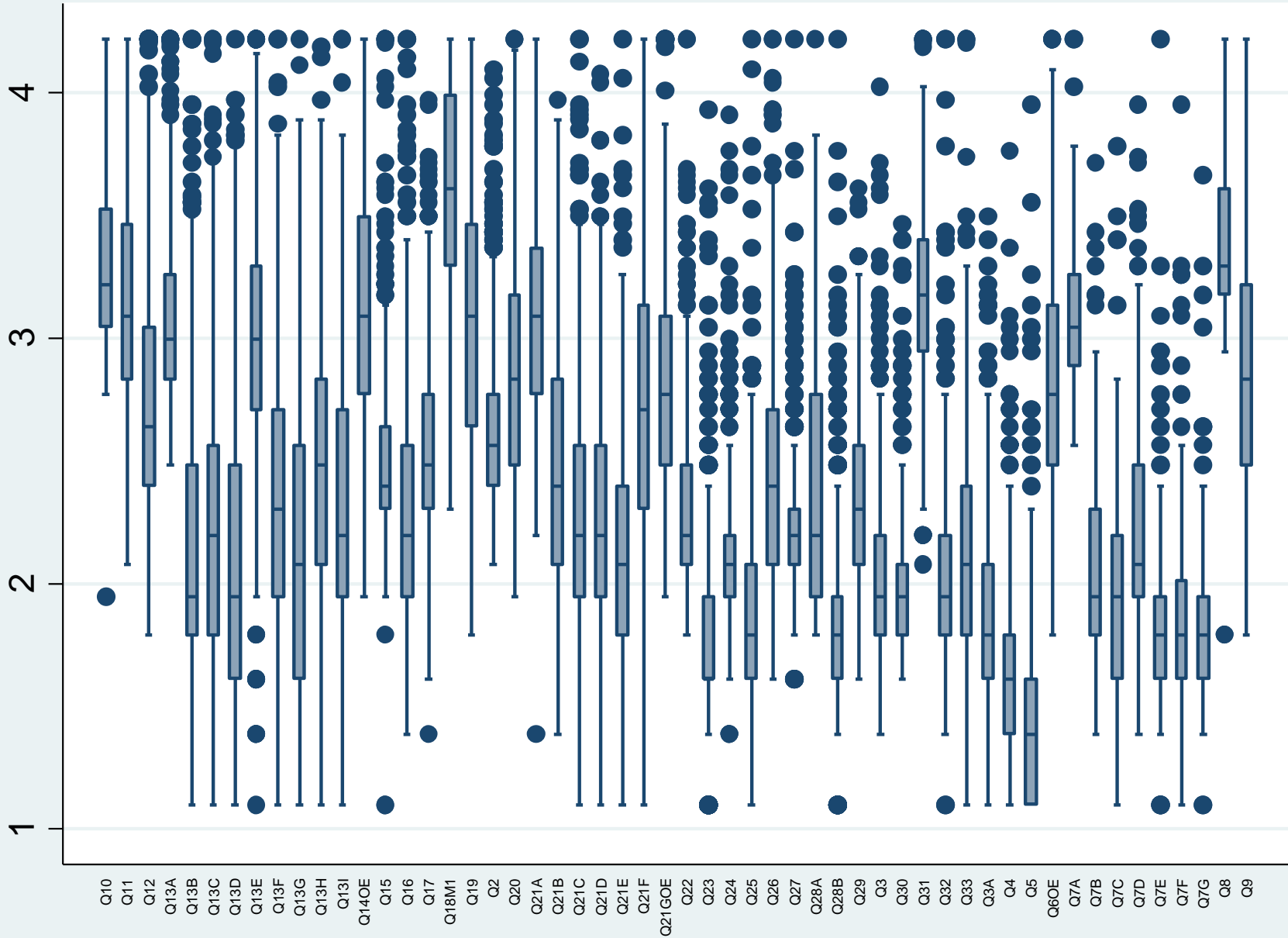
# Data

- Work and Leisure Today Survey
  - Landline RDD CATI survey
  - Conducted by AbtSRBI between July 31 and August 28, 2013
  - N=450, AAPOR RR3=6.3%
  - Average length overall = 15 minutes
  - Collected on the Voxco CATI software system



# Dependent Variables: Latency

- Number of seconds per question
  - Includes interviewer administration time, respondent answering time
- Calculated by taking the difference between start times for subsequent questions
- Log transformed and truncated to account for skewed timing data
  - Overall: 15.05 seconds (SD=13.64)
  - Log-transform: 2.45 log-seconds (SD=0.68)



# Dependent Variables: Question Reading Behaviors

Q17. Do you use the internet, at least occasionally? Include access to the internet on any mobile device, including cell phones, smartphones or tablet computers.

(DEFINITION: Occasionally means two to three times per week.)

- Do you use the internet at least occasionally? **This would** include access to the internet on any mobile device including cell phones, smart phones, or tablet computers **and occasionally would be 2 or 3 times per week.**
- **And uh** do you use the internet at least occasionally include access to the internet on any mobile device including cell phones, smart phones or tablet computers, **yes or no?**

Q10. The next question is going to ask you about how often you've engaged in exercise. We want you to include walking, anything you may do around the house, or work you do on a job if you think they constitute exercise. So using that definition, IN THE LAST WEEK, on how many days did you do any exercise?

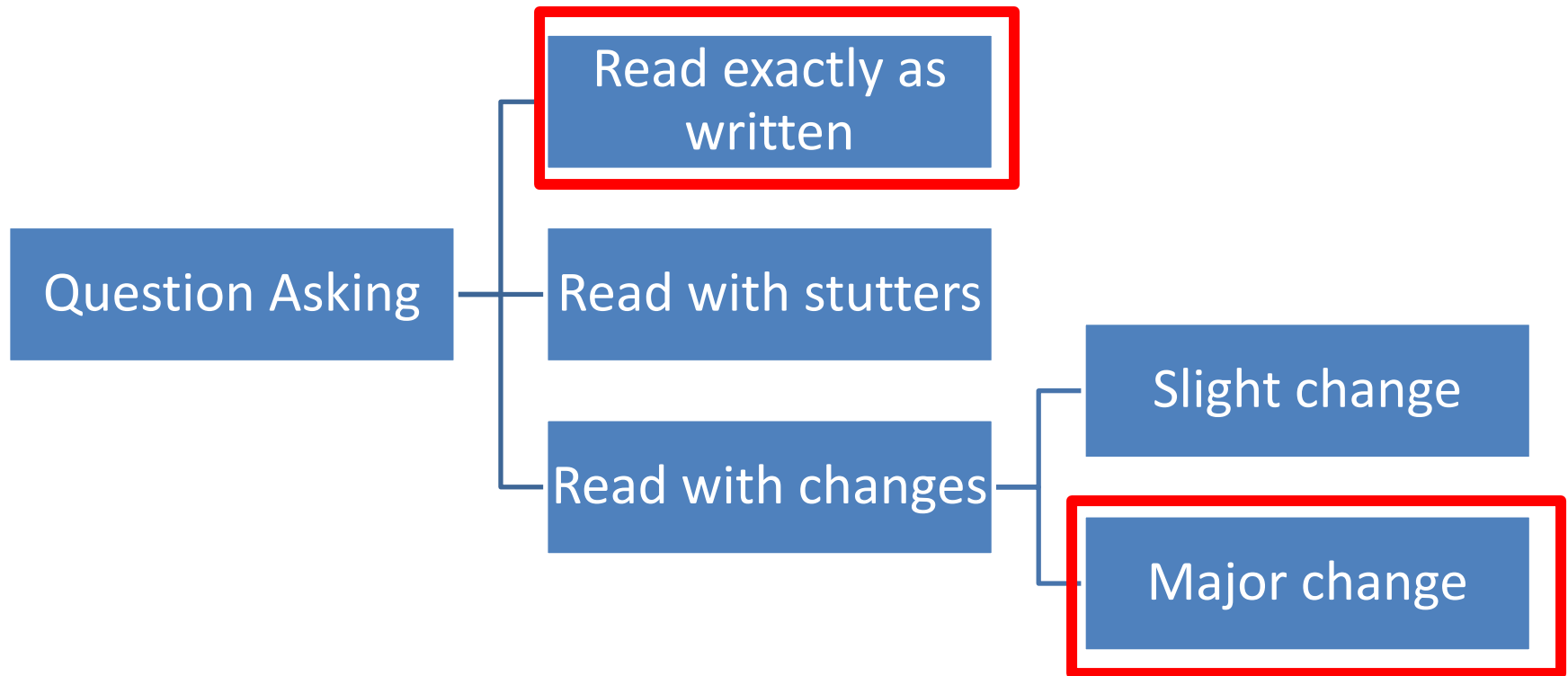
- **Now** the next question's going to ask you about how often you've engaged in exercise. We want you to include walking, anything you may do around the house or work you do on a job, **well see tha, see, this is the other population**, if you think they constitute exercise. So using that definition, in the last week how many days did you do any exercise?
- Now the next question is going to ask about how often you have engaged in **an** exercise. We want you to include walking, anything you may do around the house, or work you do on a job if you think they constitute exercise. So using that definition, in the last week, how many days did you do any exercise? **How many days in se, last week?**

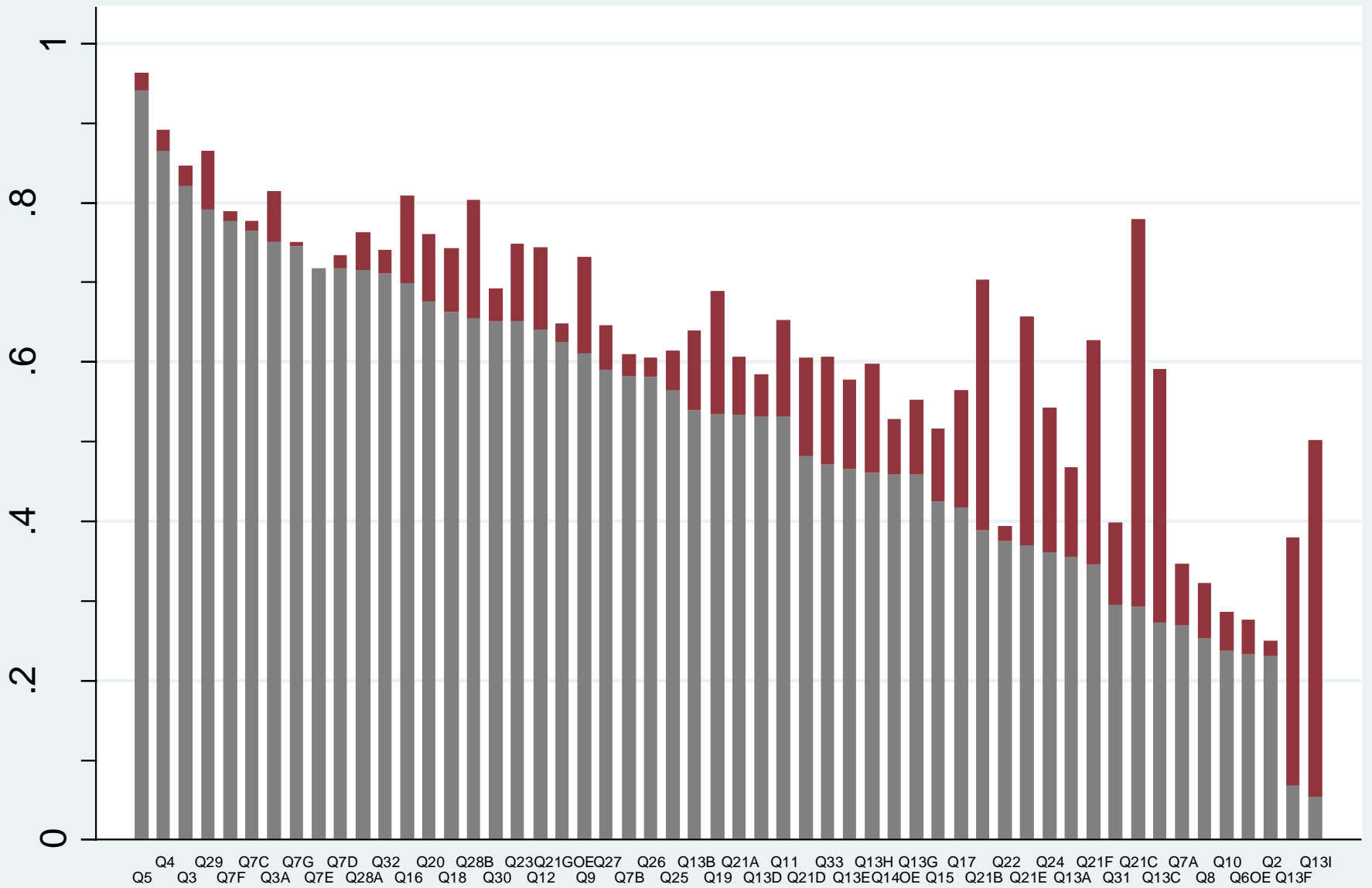
# Behavior Codes

- 8 fields coded by trained undergraduate coders
  - 10% subsample of interviews coded by two master coders

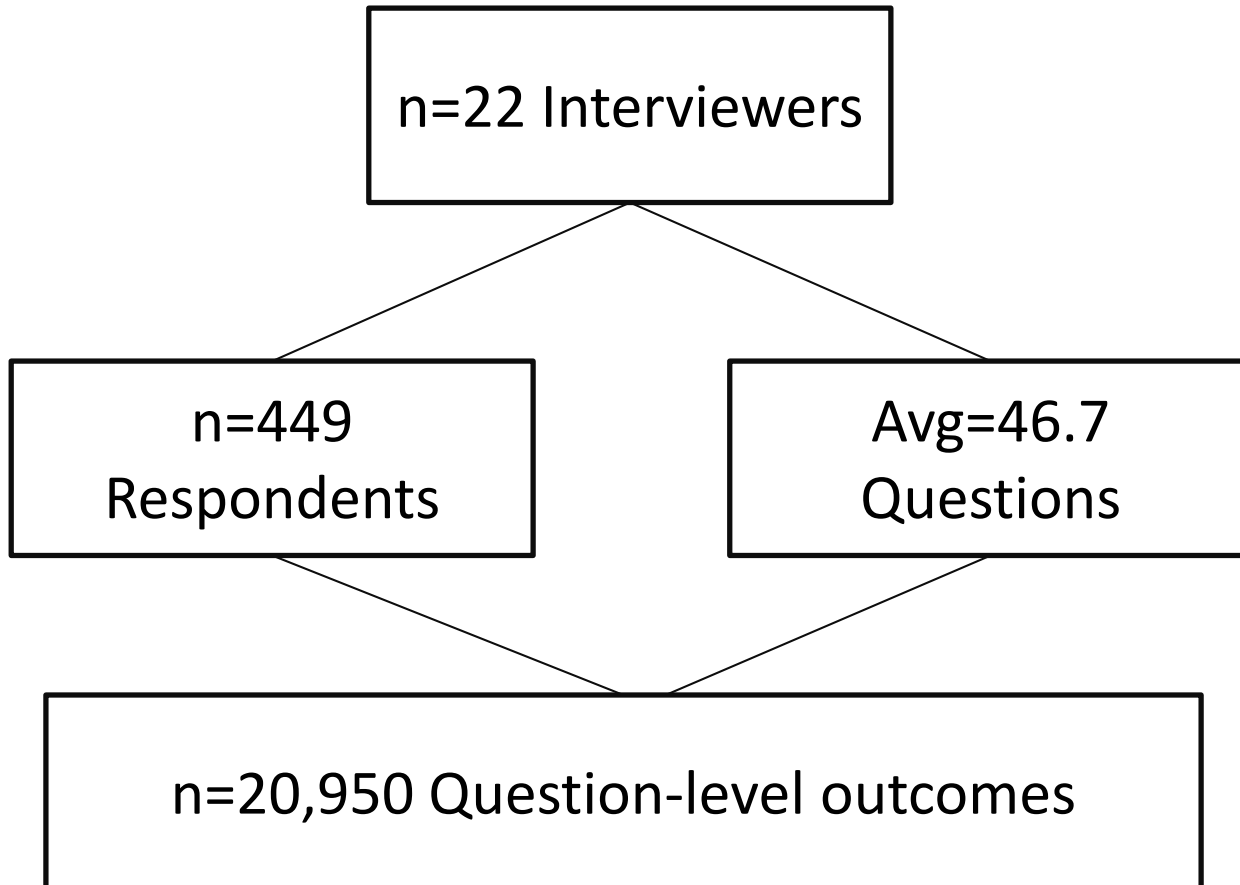
Field	Kappa
Actor	k=0.998
Initial Action	0.9
Assessment of Initial Action	0.55 to 0.68
Details of Action	0.10 to 0.77
Parentheses	0.9
Laughter	1
Disfluencies	0.9
Interruptions	0.9

# Question Reading behaviors





# Data Structure





# Modeling Approach

- Cross-classified Random Effects Models, estimated in Stata13.1 using `xtmixed` and `xtmelogit`

$$Y_{i(j_1, j_2)k} = \gamma_{0000} + \nu_{000k} + u_{0j_{10}k} + u_{00j_2k} + e_{i(j_1, j_2)k}$$

$$\text{logit}\{\text{Pr}(Y_{i(j_1, j_2)k} = 1)\} = \gamma_{0000} + \nu_{000k} + u_{0j_{10}k} + u_{00j_2k}$$

- Intraclass correlation coefficients

$$\rho_{resp} = \frac{\hat{\tau}_{uj_{10}}}{\hat{\tau}_{uj_{10}} + \hat{\tau}_{uj_2} + \hat{\tau}_{uk} + \hat{\sigma}_e^2}$$


# Model Building: Multivariate Models


- Enter each group of items for each construct jointly
- Build combined model with question characteristics
- Add respondent and interviewer characteristics
- Remove non-significant effects for parsimonious model
  - Will show only results from parsimonious model

$$Y_{i(j_1, j_2)k} = \gamma_{0000} + \sum_{m=1}^p \beta_m \text{Respondent\_char}_{j_{10}} + \sum_{s=1}^q \beta_s \text{Question\_char}_{j_2} \\ + \sum_{t=1}^r \beta_t \text{Iwer\_char}_k + v_{000k} + u_{0j_{10}k} + u_{00j_2k} + e_{i(j_1, j_2)k}$$

# Independent variables

- Double coded each screen in WLT survey
  - Kappas range from 0.85 to 1.00

Progress  5

Navigate : Q8      LANG

Q8

Q8. We are interested in volunteer activities for which people are not paid, except perhaps expenses. We only want you to include volunteer activities that you did through or for an organization, even if you only did them once in a while. In the last 12 months, that is since July of last year (2012), have you done any volunteer activities through or for an organization?

[PROBE: IF HAVE NOT VOLUNTEERED, ASK:] Sometimes people don't think of activities they do infrequently or activities they do for children's schools or youth organizations as volunteer activities. Since July of last year, have you done any of these types of volunteer activities?

1  Yes  
2  No  
8  (VOL) Don't know  
9  (VOL) Refused

# Question Characteristics

- Necessary Question characteristics
  - Mean # words in question = 14.75 (SD=12.92)
  - Mean # response options = 3.42 (SD=3.55)
  - 28% Attitude, 46% Behavior, 27% Demographics
  - 10% Open-ended text, 34% open-ended numeric, 27% closed-ended nominal (including yes/no), 30% closed-ended ordinal
  - 52 questions in questionnaire (including all skip patterns)
- Respondent Task Complexity
  - Mean reading level = 6.71 (SD=4.83)
  - 14% mismatch between question and response options
  - 15% sensitive questions
  - 4% unknown terms

# Question Characteristics

- Visual Design
  - 40% interviewer instructions
  - 11% parentheses
  - 34% question asked over two screens
  - 6% probes on screen
  - 14% visual emphasis
- Processing efficiency
  - 7% first question in battery; 31% later questions in battery
  - 19% definitions
  - 13% transition statement
  - 6% feeder question; 20% later questions

# Question characteristics: Challenges

- Question characteristics are not randomly assigned
  - Question characteristics appear in a “package” – and are thus multicollinear.
  - Examples:
    - Questions with transition statements have more words (PBS correlation=0.48).
    - Demographic questions are (largely) asked at the end of the survey.

# Respondent characteristics

- Mean age = 61.4 years (SD=16.71)
- 71% Associate's degree +
- 64% Female
- 78% Computer at home
- 41% Employed

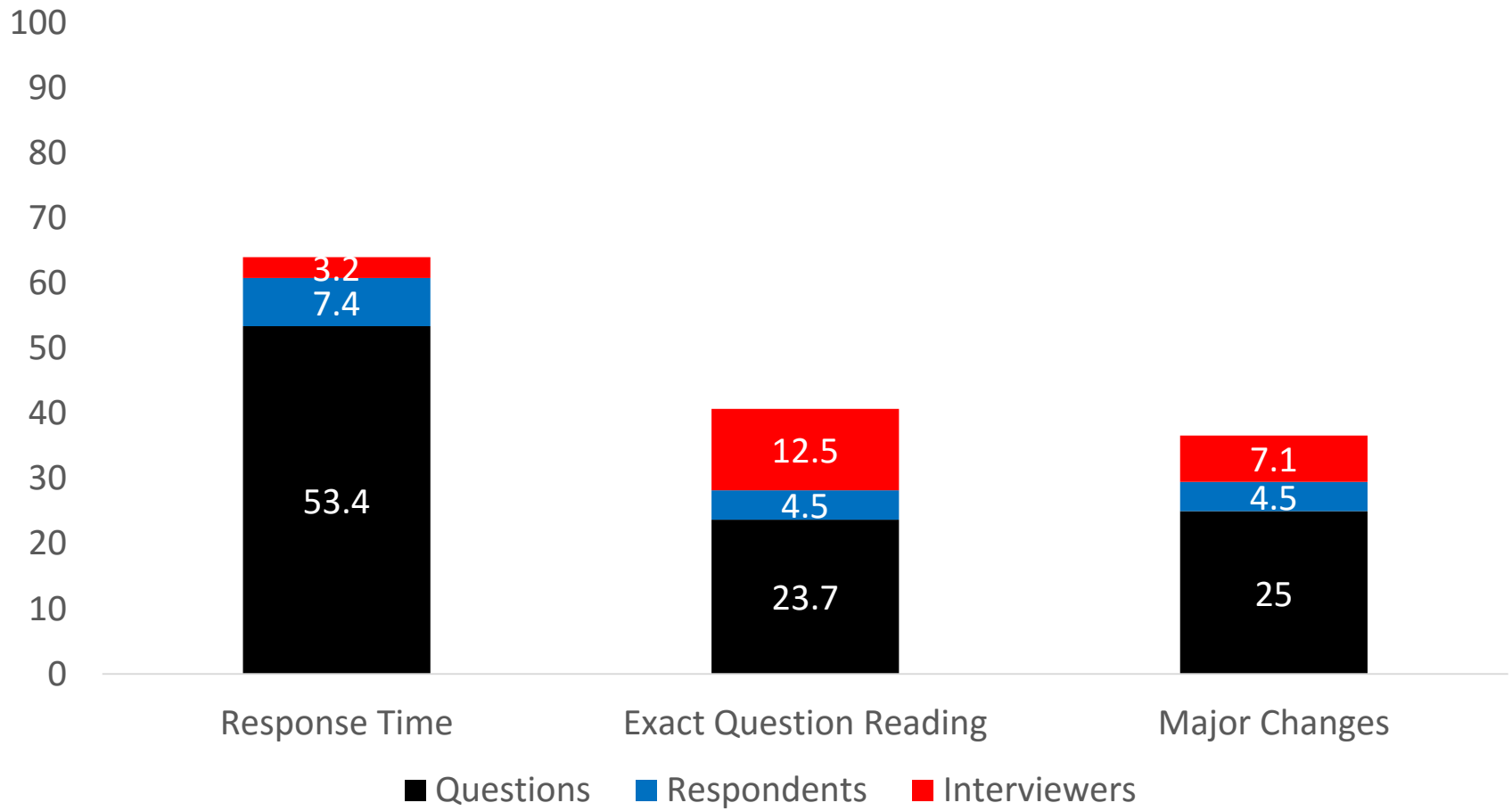
# Interviewer characteristics

- 68% 1+ Years of experience
- 54% Female
- 59% White
- Mean Cooperation rate = 6.7% (SD=2.0)



# Findings

# Findings: Base Model, Intraclass correlation coefficients



# Findings: necessary question features

	Response timing
Ordering of questions	--
Length of question	0.020****
# of response options	0.041****
Type of question	
Behavior (ref)	--
Attitude/opinion	-1.080***
Demographics	-0.292**
Format of response options	
Open-ended text (ref)	--
Open-ended numeric	-0.310*
Closed-nominal	-0.626****
Closed-ordinal	0.298
Yes-No	-0.844****

# Necessary question characteristics

	Response Latencies	Exact reading	Major Misreading
<b>Ordering of questions in questionnaire</b> (e.g., Couper and Kreuter 2013; Yan and Tourangeau 2008; but Holbrook et al. 2015)	↓	↓	↑
<b>Length of question</b> (e.g., Yan & Tourangeau 2008; Childs and Landreth 2006; Ongena and Dijkstra 2007)	↑	↓	↑
<b># of response options</b> (e.g, Yan & Tourangeau 2008, but Krosnick & Fabrigar 1997; Fowler and Mangione 1990, Childs and Landreth 2006; but Holbrook et al. 2015)	↑	?	?

# Necessary question characteristics

	Response Latencies	Exact Reading	Major Misreading
<p>Type of question (e.g., Bassili &amp; Fletcher 1991; Tourangeau, Rips, &amp; Rasinski. 2000; Yan &amp; Tourangeau 2008)</p> <ul style="list-style-type: none"> <li>• Attitude</li> <li>• Behavior</li> <li>• Demographics</li> </ul>	<p>↓</p> <p>↑</p> <p>↓</p>	<p>?</p> <p>?</p> <p>↑ ↓</p>	<p>?</p> <p>?</p> <p>↑ ↓</p>
<p>Format of response options (e.g., Smyth, et al. 2009; Couper, et al. 2011; Couper and Kreuter 2013; Childs and Landreth 2006; Mathiowetz and Cannell 1980)</p> <ul style="list-style-type: none"> <li>• Open-ended (numeric vs. text)</li> <li>• Closed-ended nominal</li> <li>• Closed-ended ordinal</li> <li>• Yes-no</li> </ul>	<p>↓</p> <p>↑</p> <p>↓</p> <p>↓</p>	<p>--</p> <p>--</p> <p>↑</p> <p>↑</p>	<p>--</p> <p>--</p> <p>?</p> <p>?</p>

# Findings: respondent task complexity

	Response timing	
Question reading level	0.027****	
Mismatch between question and response options	--	
Unknown terms	--	
Sensitive questions	-0.337**	

# Respondent task complexity

	Response Latencies	Exact reading	Major Misreading
Question reading level (e.g., Yan & Tourangeau 2008; Fowler and Mangione 1990; Ongena and Dijkstra 2007)	↑	↓	↑
Mismatch between question and response options (e.g., Dykema, Schaeffer, Garbarski, 2016)	↑	↓	↑
Unknown terms (Fowler, 1992)	↑	?	?
Sensitive questions (e.g. Ongena and Dijkstra 2007; but Fowler and Mangione 1990; Holbrook et al. 2015)	↓	?	?

# Findings: Processing efficiency

	Response timing	
First question in battery		
Later questions in battery	--	
Questions with definitions	0.212*	
Transition statement	--	
Feeder question	--	
Follow-up question in a skip	--	



# Processing efficiency

	Response Latencies	Exact reading	Major Misreading
First question in battery (Dykema, Schaeffer and Garbarski, 2016)	↑	↓	↑
Later question in battery (Dykema, Schaeffer and Garbarski, 2016)	↓	↓	↑
Questions with definitions (Fowler, 1992)	↑	↓	↑
Transition statement	↓	↓	↑
Feeder question	--	?	?
Follow-up question in a skip	↓	--	--

# Findings: Visual design of screen

	Response timing	
Interviewer instructions	--	
Probe on screen	--	
Parentheses in question stem	--	
Question asked over two screens	--	
Visual emphasis	--	

# Visual Design of Screen

	Response Latencies	Exact reading	Major Misreading
Interviewer instructions (but see Couper and Kreuter 2013)	↑	?	?
Probe on screen	↑	↓	↑
Use of parentheses in question stem	↑	↓	↑
Questions asked over two screens	↑	--	--
Questions with visual emphasis	↑	--	--

# Respondent and interviewer characteristics

## Respondent

- Longer time and fewer exact readings for older respondents
- Shorter time and more exact readings for employed respondents and respondents with computers

## Interviewer

- No effects of any interviewer characteristics

# Variance explained

- Of the initial variability due to questions:
  - Response time: 82% explained
  - Exact question reading: 75% explained
  - Major question misreading: 57% explained
- Little of the variance due to respondents or interviewers explained with the covariates presented here.

# Summary: Response time

- Survey logistics and respondent task complexity measures are primary predictors of response time
- Replicated effects from existing research
  - Also found that sensitive questions are faster and questions with definitions are longer
- Little of the variation in response time is attributable to interviewers or respondents

# Summary: Question reading

- Question-level predictors of reading questions correctly and making major errors in reading differ.
  - But both occur differentially for questions with parentheses in the question stem and in battery items.
    - Neither of these question characteristics are associated with response time!
  - Both also change over the course of the questionnaire, net of the characteristics of the questions.
  - Other visual design (interviewer probes) and question characteristics (length of question, type of question) predict reading questions exactly as written, but not major misreadings.
- More of the variability in question reading behaviors is due to interviewers, as expected.
  - But not able to predict this variability with interviewer characteristics included.

# What we learned

- Theoretical mechanisms behind response time largely respondent-driven, not interviewer driven
  - Question reading behaviors make up about 55% of the total time spent on a question, on average.
  - Evidence of some adaptation of interviewers to different types of respondents.
- Paradata-derived response times are a poor indicator of the types of questions that are read poorly.
  - If the goal of using paradata as a diagnostic tool is to identify reading problems, response time won't do it.



# Implications for survey design

- Much of the determinants of response timing are directly determined by the content of the question, not visual design
- Visual design has a larger effect on question reading behaviors.
  - But many of the correlates of poor question reading don't affect overall response time.
- Is “faster” or “slower” better? Do questions read as written generate better data quality?
  - Our (current) findings don't allow us to evaluate this

# Limitations

- Interviewers were new to Voxco
- Landline sample
- No experimental assignment of question characteristics
  
- All of these limitations addressed in the Work and Leisure Today 2 Study fielded during summer 2015
  - Behavior coding is underway

# Future research

- Differences in effects of question characteristics over respondents
- Respondent behaviors (Ganshert, et al. 2016 AAPOR)
- Differences for cell and landline telephones (Timbrook, et al. 2016 AAPOR)
- What is happening in battery questions? (Cochran, et al. 2016 AAPOR)

For more information on response  
time, see

- 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. doi:10.1093/jssam/smv021

Thank you!

[kolson5@unl.edu](mailto:kolson5@unl.edu)