2-26-2019

The Accuracy and Utility of Using Paradata to Detect Interviewer Question-Reading Deviations

Jennifer Kelley

The Accuracy and Utility of Using Paradata to Detect Interviewer Question-Reading Deviations

Jennifer Kelley, University of Essex

Interviewer Workshop, University of Nebraska-Lincoln, 2019

Presentation Outline

- Motivation for Research
- Background
- Data and Methods
- Results
- Conclusion

Motivations for Research

- Interviewers' behavior at training vs. behavior in field

Background

- Interviewers and measurement error

- How to reduce measurement error?
  - Training interviewers to read questions verbatim
  - Supervising and monitoring interviewers

- Do interviewers read question verbatim?
  - Studies show question-reading deviations range from 4.6% - 84.0%

Monitoring Interviewer Question-reading Behavior

- Listen to interview recordings

Monitoring Interviewer Behavior with Paradata

- Timestamp is as a proxy for how the interviewer reads the question
- Estimate how long it **should** take interviewers to read a question
- Create ***question administration timing threshold*** (QATT)
- Compare the QATT to the question timestamp
- Known studies that use timestamps and QATTs
  - Saudi National Mental Health Survey
    - Flagged questions that have timestamps under 1 second
  - China Mental Health Survey
    - Calculated QATT using the number of words in the question and reading pace of 110 millisecond per Chinese Character

Advantages of Using Timestamps to Monitor Question-reading Behavior

- Automate process
- Fast
- Target QC efforts

Present Study

- **Accuracy and utility of method currently used?**
- **More accurate method for developing QATTs?**
  - WPS Range
  - Standard deviation
  - Model-based
    - Study attempts to identify ‘cheating’ in web-surveys (Munzert & Selb, 2015)
    - Latency as indicator for potential cheating
    - Response times are mostly likely both person and item specific
    - Model response times as a function of person specific random intercepts and fixed effects for items specific factors to isolate “suspicious latency”
    - Extracted residuals and classified top 2% as cheaters

Data

- Wave 3 of the Understanding Society Innovation Panel
  - Multi-stage probability sample
- 1621 CAPI interviews
- Interviewers are trained to read all questions verbatim
- Sections of the interview were recorded with permission of respondent
- Interview recordings
  - 820 recordings were available for analysis
  - Interviewers were told which sections would be recorded
- Paradata: timestamps for all questions across all interviews

Methods

- Randomly selected two recorded interviews from each interviewer (n=81) and behavior coded all selected questions in the recording
- Selected questions based on following criteria
  - Question was intended to be read out loud
  - Did not contain ‘fills’
  - Were administered to both males and females
  - Had one-to-one matching with timing file questions (i.e., did not loop)
  - Had same response options for all regions
- Total sample size: 10,345 questions

Methods: Behavior Coding

- Interviewer's first reading of the question was coded
- Verbatim or Deviation
- Magnitude of deviation
  - Minor
  - Major

More Details on Behavior Coding

- Deviations were coded as major deviations under any of the following circumstances:
  - Key nouns, verbs or adjectives/qualifiers were omitted
  - Key nouns, verbs or adjectives/qualifiers were subbed with words that did not have equivalence in meaning
  - Key nouns, verbs or adjectives/qualifiers were added that altered the context or added additional (inaccurate) meaning
  - Definitions or examples were omitted that were needed to give context to the question
  - Definitions or examples were subbed with words that did not retain equivalence in meaning
  - Unfamiliar response options were omitted that were needed to ensure all respondents were received same range of options (e.g., "Do you work for *a private firm or business or other limited company or do you work for some other type of organization*?")

Methods: Constructing QATTs

- Minimum QATTs based on words per second
  - 2wps, 3wps, 4wps

- Minimum and maximum QATTS based on
  - Range WPS
    - 2-3wps, 2-4wps, 1-3wps, 1-4wps

Methods: Constructing QATTs

- Standard deviation
  - ±0.5 SD, ± 1 SD, ± 1.5 SD, ± 2.0 SD
- Model-based
  - Timestamps (logged) to each question are predicted by a model with random intercept for interviewer and fixed effects for the respondent and question ID
  - Residuals standardized into a t-score and categorized the upper and lower t-distribution as possible deviations
  - 1%, 2%, 3%, 5%, 10%, and 25%

Methods: Variables and Analysis

- Detection method variable
  - Question timestamp compared to the question QATT for each detection method
  - 0=Verbatim, 1=Deviation
- Behavior coding variable
  - 0=Verbatim, 1=Minor deviation, 2=Major deviation
- Crosstabs to determine accuracy of each detection method
  - Produces rates for

    X **False –** (incorrectly identified deviation as verbatim)

    X **False +** (incorrectly identified verbatim as deviation)

    ✓**True –** (correctly identified verbatim as verbatim)

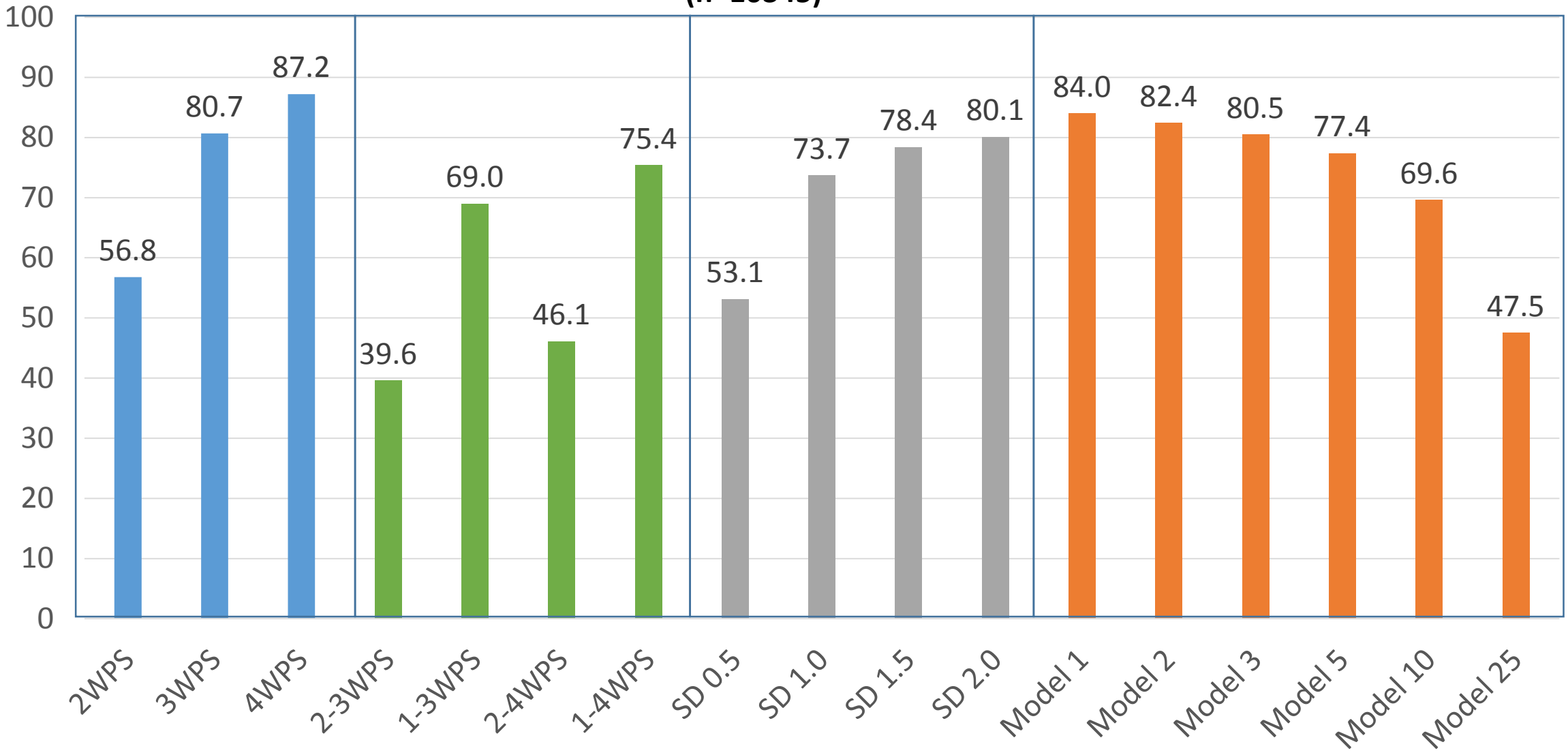
    ✓**True +** (correctly identified deviation as deviation)

What Does the Behavior Coding Tell Us?

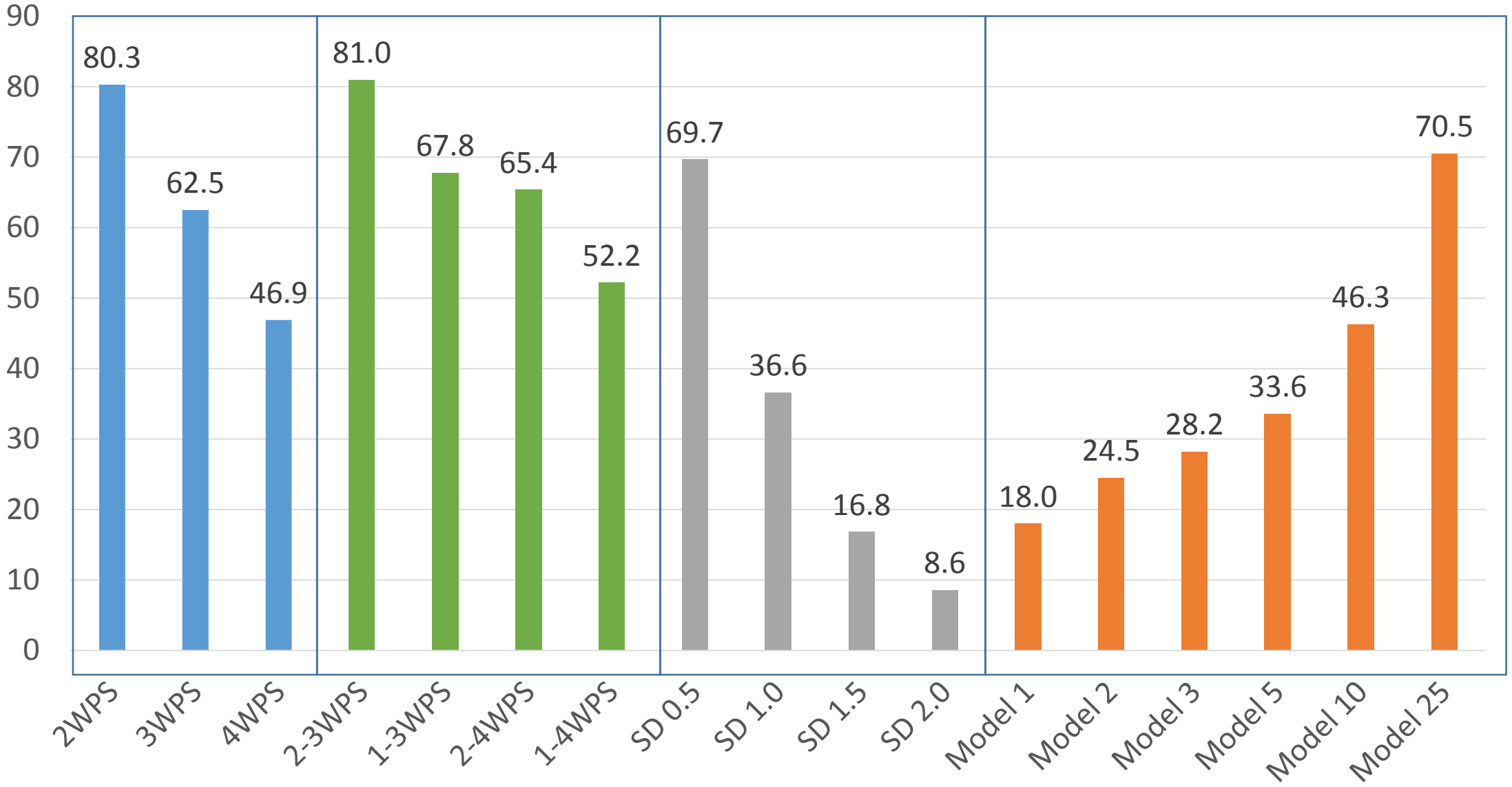
Question Reading (n=10345)	Count	
Verbatim	5435	52.5
Minor Deviation	3567	34.5
Major Deviation	1343	13.0

Accuracy Rate (%) for Correctly Identifying Questions as Major Deviations and No Major Deviation (i.e. verbatim/minor)

(n=10345)



Detection Rate (%) for Correctly Identifying Major Deviations (n=1343)



Accuracy Rate (%) of Detecting Deviations: QATT Detection Methods by Major Deviation (n=10345)

	Overall Accuracy	Detection Rate	False -	False +	True -	True +
4WPS	87.2	46.9	6.9	6.0	81.1	6.1
2-3WPS	39.6	81.0	2.5	57.9	29.1	10.5

Utility of the QATT Methods

- False positive and false negatives may be reduced if the data is aggregated up to the interview level
- Data was aggregated to the interview level (n=168)
- All interviews contained at least one minor deviation and 139 (82.7%) of interviews contained at least one major deviation
- Which method is best at reducing QC efforts, but still identifies all interviews that contain at least one major deviation?

Interview Level Analysis

- Some methods correctly flagged all interviews that contained at least one major deviation…..but flagged all interviews for review

- 4WPS shows promise
  - Correctly flagged 132 of the 139 interviews that contained at least one major deviation
  - Correctly flagged 17 or the 29 interviews with no major deviations
  - 85.7% of interviews flagged for review

Discussion: Summary

- As overall accuracy increases, false negatives also increase
- As detection rate increases, false positives also increase
- 4WPS has the highest overall accuracy rate - 87.1%, but only detects 46.9% of the major deviations
- 2-3WPS method is best at detecting potential major deviations 81.0%, but produces the highest rate of false positives – 57.9%
- 4WPS shows the most utility at the interview level
- WPS range, SD, and model-based methods did not do as well as the WPS Method

- Special Thanks
  - Tarek Al Baghal, Supervisor
  - Peter Lynn, Supervisor

Thank you! Feedback is welcomed and appreciated!

Contact info: jennifer.kelley@essex.ac.uk

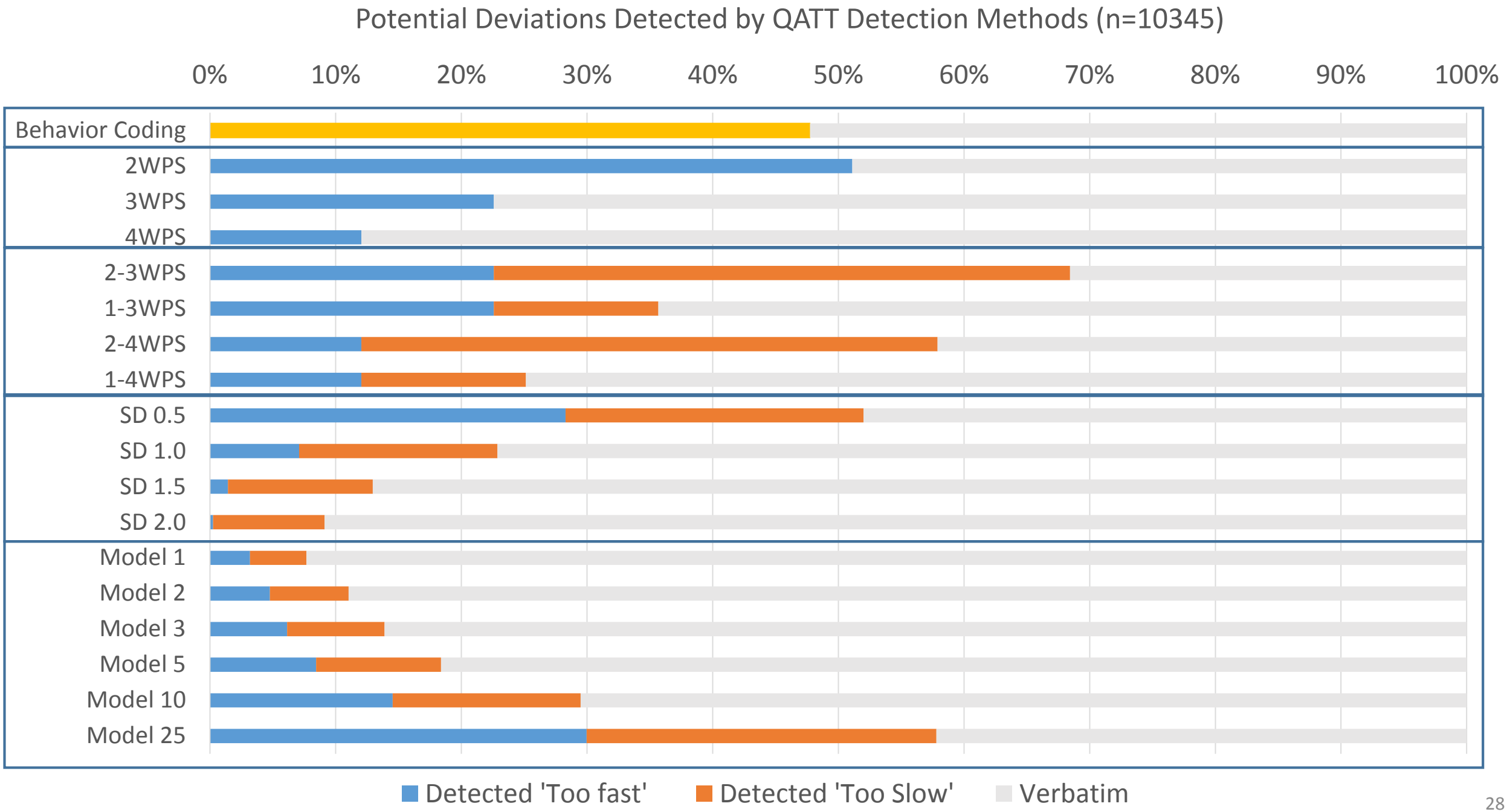
Additional Slides for Discussion

Future Research

- Second Paper: What drives question-reading deviations?
  - Question, respondent and interviewer characteristics
- Third Paper: Data quality
  - So interviewers make deviations from reading verbatim – does it mater?
- Accuracy and Utility 2.0
  - Test different models
  - Use data from previous waves to create QATTs
  - Use paradata files that have timestamps in milliseconds rather than seconds
- Can timestamps and QATTs be used for methodological research?

More Details on Behavior Coding

- Deviations were coded as minor deviations under the following circumstances :
  - Omitted, subbed or added articles (e.g., the, an, this, etc.)
  - Omitted or subbed a time reference (e.g., “Since we last interviewed you [omit: on January 22, 2008] did…”)
  - Interview instructions omitted, subbed or added that did not give meaning or context to question (e.g., please look at the card)
  - Interviewer omitted response options starting on the second question of a series of questions (e.g., always, very often, quite often, not very often, never)
  - Respondent interrupted the interviewer to signal their correct response for previously heard response options (e.g., agree, neither agree nor disagree, disagree)
  - Skipped the entire question, but response was given in previous answer
- Deviations were coded as major deviations under any of the following circumstances:
  - Key nouns, verbs or adjectives/qualifiers were omitted
  - Key nouns, verbs or adjectives/qualifiers were subbed with words that did not have equivalence in meaning
  - Key nouns, verbs or adjectives/qualifiers were added that altered the context or added additional (inaccurate) meaning
  - Definitions or examples were omitted that were needed to give context to the question
  - Definitions or examples were subbed with words that did not retain equivalence in meaning
  - Non-common response options were omitted that were needed to give context to the question to ensure all respondents were received same range of options (e.g., “Do you work for *a private firm or business or other limited company or do you work for some other type of organization*?”)


Potential Deviations Detected by QATT Detection Methods (n=10345)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Behavior Coding
2WPS
3WPS
4WPS
2-3WPS
1-3WPS
2-4WPS
1-4WPS
SD 0.5
SD 1.0
SD 1.5
SD 2.0
Model 1
Model 2
Model 3
Model 5
Model 10
Model 25
Detected 'Too fast'
Detected 'Too Slow'
Verbatim

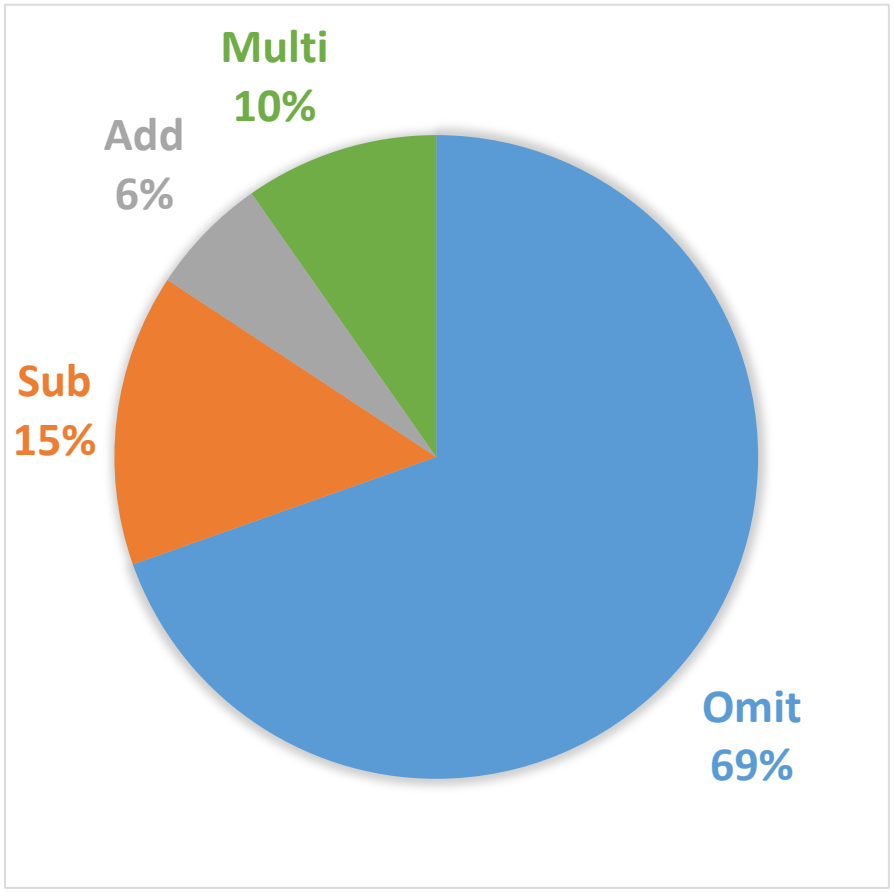
Accuracy Rate (%) of Detecting Deviations: QATT Detection Methods by Major Deviation (n=10345)

	Detected 'Too fast'					Detected 'Too slow'					Total Deviations Detected				
	False -	False +	True -	True +	Overall Acc	False -	False +	True -	True +	Overall Acc	False -	False +	True -	True +	Overall Acc
2WPS	2.6	40.7	46.3	10.4	56.8						2.6	40.7	46.3	10.4	56.8
3WPS	4.9	14.5	72.6	8.1	80.7						4.9	14.5	72.6	8.1	80.7
4WPS	6.9	6.0	81.1	6.1	87.2						6.9	6.0	81.1	6.1	87.2
2-3WPS	4.9	14.5	72.6	8.1	80.7	10.6	43.5	43.6	2.4	46.0	2.5	57.9	29.1	10.5	39.6
1-3WPS	4.9	14.5	72.6	8.1	80.7	12.3	12.4	74.6	0.7	75.3	4.2	26.9	60.1	8.8	69.0
2-4WPS	6.9	6.0	81.1	6.1	87.2	10.6	43.5	43.6	2.4	46.0	4.5	49.4	37.6	8.5	46.1
1-4WPS	6.9	6.0	81.1	6.1	87.2	12.3	12.4	74.6	0.7	75.3	6.2	18.4	68.7	6.8	75.4
SD 0.5	5.9	21.2	65.8	7.1	72.9	11.0	21.7	65.3	2.0	67.3	3.9	42.9	44.1	9.0	53.1
SD 1.0	9.7	3.8	83.2	3.3	86.5	11.5	14.3	72.7	1.5	74.2	8.2	18.1	68.9	4.8	73.7
SD 1.5	12.0	0.4	86.6	1.0	87.5	11.8	10.3	76.7	1.2	77.9	10.8	10.8	76.2	2.2	78.4
SD 2.0	12.8	0.0	87.0	0.2	87.2	12.1	8.0	79.1	0.9	80.0	11.9	8.0	79.0	1.1	80.1
Model 1	11.3	1.5	85.5	1.7	87.2	12.3	3.8	83.2	0.7	83.9	10.6	5.3	81.7	2.3	84.0
Model 2	10.7	2.5	84.5	2.3	86.8	12.1	5.4	81.7	0.9	82.6	9.8	7.8	79.2	3.2	82.4
Model 3	10.3	3.5	83.5	2.6	86.2	12.0	6.7	80.3	1.0	81.3	9.3	10.2	76.8	3.7	80.5
Model 5	9.8	5.3	81.7	3.1	84.9	11.8	8.7	78.3	1.2	79.5	8.6	14.0	73.0	4.4	77.4
Model 10	8.8	10.3	76.7	4.2	80.9	11.2	13.1	73.9	1.8	75.7	7.0	23.5	63.5	6.0	69.6
Model 25	6.9	23.9	63.1	6.1	69.2	9.9	24.8	62.2	3.1	65.3	3.8	48.7	38.4	9.2	47.5

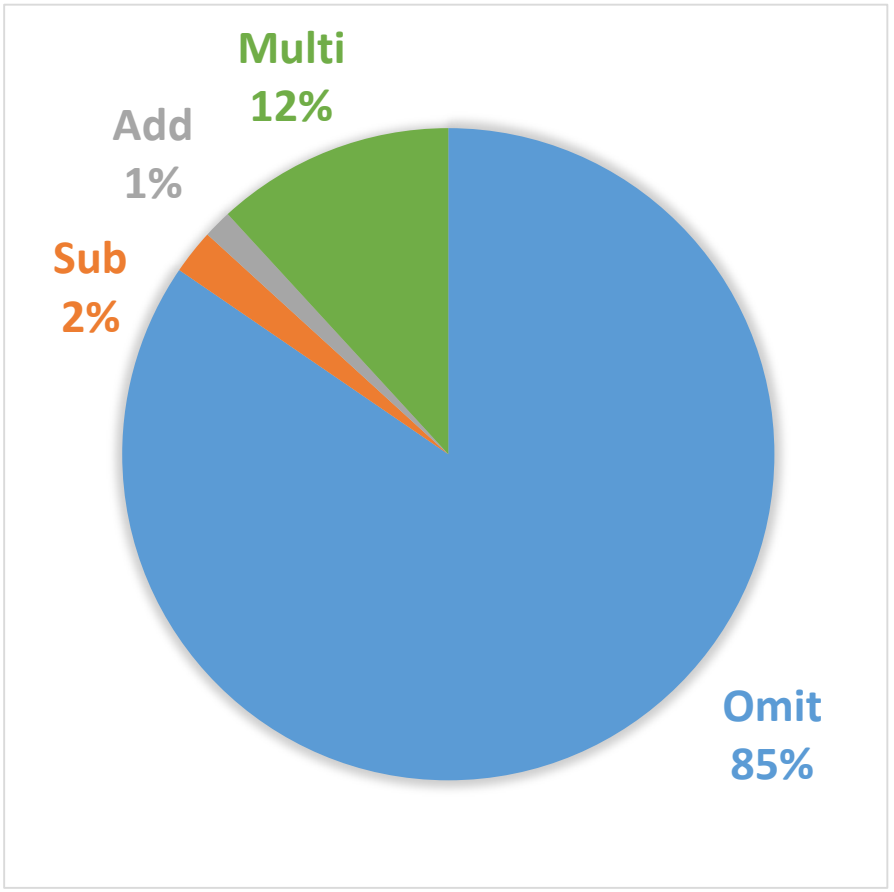
Detection Method	Count of Interviews Correctly Flagged As Containing:		Count of Interviews Incorrectly Flagged as Containing:		Overall Accuracy (%)	% of Interviews Deviation Detected n=139	Interviews Method Flagged for Review (%)
	Deviation	No Deviation	Deviation	No Deviation			
2WPS	139	0	29	0	82.7	100.0	100.0
3WPS	137	6	23	2	85.1	98.6	95.2
4WPS	132	17	7	12	88.7	95.0	82.7
2-3WPS	139	0	29	0	82.7	100.0	100.0
1-3WPS	139	0	29	0	82.7	100.0	100.0
2-4WPS	139	0	29	0	82.7	100.0	100.0
1-4WPS	138	4	25	1	84.5	99.3	97.0
SD 0.5	139	0	29	0	82.7	100.0	100.0
SD 1.0	139	3	26	0	84.5	100.0	98.2
SD 1.5	134	10	19	5	85.7	96.4	91.1
SD 2.0	124	13	16	15	81.5	89.2	83.3
Model 1	127	6	23	12	79.2	91.4	89.3
Model 2	133	2	27	6	80.4	95.7	95.2
Model 3	137	2	27	2	82.7	98.6	97.6
Model 5	139	1	28	0	83.3	100.0	99.4
Model 10	139	0	29	0	82.7	100.0	100.0
Model 25	139	0	29	0	82.7	100.0	100.0

Behavior Coding: Types of Deviations

Minor Deviations (n=3567)



Major Deviations (n=1343)



References

Ackermann-Piek, D., & Massing, N. (2014). Interviewer behavior and interviewer characteristics in PIAAC Germany. Methods, data, analyses: a journal for quantitative methods and survey methodology (mda), 8(2), 199-222.

Axinn, W. G. (1991). The influence of interviewer sex on responses to sensitive questions in Nepal. Social Science Research, 20(3), 303-318.

Bassili, J. N. (1996) The how and the why of response latency measurement in telephone surveys. In Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research (eds N. Schwarz and S. Sudman), pp. 319–346. San Francisco: Jossey-Bass.

Bassili, J. N. and Fletcher, J. F. (1991). Response-Time Measurement in Survey Research a Method for CATI and a New Look at Nonattitudes. Public Opinion Quarterly, 55(3): 331-346.

Cannell, C. F. (1975). A Technique for Evaluating Interviewer Performance.

Conrad, F. G., Broome, J. S., Benkí, J. R., Kreuter, F., Groves, R. M., Vannette, D., & McClain, C. (2013). Interviewer speech and the success of survey invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 191-210.

Couper, M. P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments.
Social Science Computer Review, 18(4):384-396.

Draisma, S. and Dijkstra, W. (2004) Response latency and (para) linguistic expressions as indicators of response error. In Methods for Testing and Evaluating Survey Questionnaires (eds S. Presser, J. Rogthgeb, M. Couper, J. Lessler, E. Martin, J. Martin and E. Singer), pp. 131–147. Hoboken: Wiley.

Fowler Jr, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions.

Groves, Robert M., et al. Survey methodology. Vol. 561. John Wiley & Sons, 2011.

Jans, M., Sirkis, R., & Morgan, D. (2013). Managing Data Quality Indicators with Paradata Based Statistical Quality Control Tools: The Keys to Survey Performance. Improving Surveys with Paradata, 191-229.

Kirgis, N., et al. (2015). Using paradata to monitor interviewer behavior and reduce survey error. TSE.

Kreuter, F. (2013). Improving surveys with paradata: Introduction. *Improving Surveys with Paradata*, 1-9.

References (cont.)

Krosnick, J. A., Malhotra, N., & Mittal, U. (2014). Public misunderstanding of political facts: How question wording affected estimates of partisan differences in birtherism

Munzert, S., & Selb, P. (2015). Measuring Political Knowledge in Web-Based Surveys: An Experimental Validation of Visual Versus Verbal Instruments. Social Science Computer Review, 0894439315616325.

Mneimneh, Z. N., Pennell, B., Lin, Y., & Kelley, J. (2014). Using paradata to monitor interviewers' behavior: A case study from a national survey in the Kingdom of Saudi Arabia. Comparative Survey Design and Implementation (CSDI) conference

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations.

Omoigui, N., He, L., Gupta A., Grudin, J. and Sanocki, E. (1999), Time-compression: Systems concerns, usage, and benefits, CHI 99 Conference Proceedings, 136-143.

Ongena, Y. P., & Dijkstra, W. (2006). Question-answer sequences in survey-interviews. Quality & Quantity, 40, 983-1011. doi: 10.1007/s11135-005-5076-4

Rugg, D. (1941). Experiments in wording questions: II. Public Opinion Quarterly, 5(1), 91.

Schober, M. F., & Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer & J. van der Zouven (Eds.), Standardization and tacit knowledge: interaction and practice in the survey interview (pp. 67-94). New York, NY: John Wiley & Sons.

Schuman, H., & Presser, S. (1996). Questions and answers in attitude surveys: Experiments on question form, wording, and context. Sage.

Sun, Y., & Meng, X. (2014). Using response time for each question in quality control on China Mental Health Survey (CMHS). Comparative Survey Design and Implementation (CSDI) conference

Wagner, J. (2013). Using Paradata-Driven Models to Improve Contact Rates in Telephone and Face-to-Face Surveys. Improving Surveys with Paradata, 145-1

West, B. T., & Sinibaldi, J. (2013). The quality of paradata: A literature review. *Improving Surveys with Paradata*, 339-359.

Yan, T. and Tourangeau, R. (2008) Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Appl. Cogn. Psychol., 22, 51–68.