Time Is Money: Methods to Measure and Reduce Survey Burden

Bryan Wu, Kaiser Family Foundation

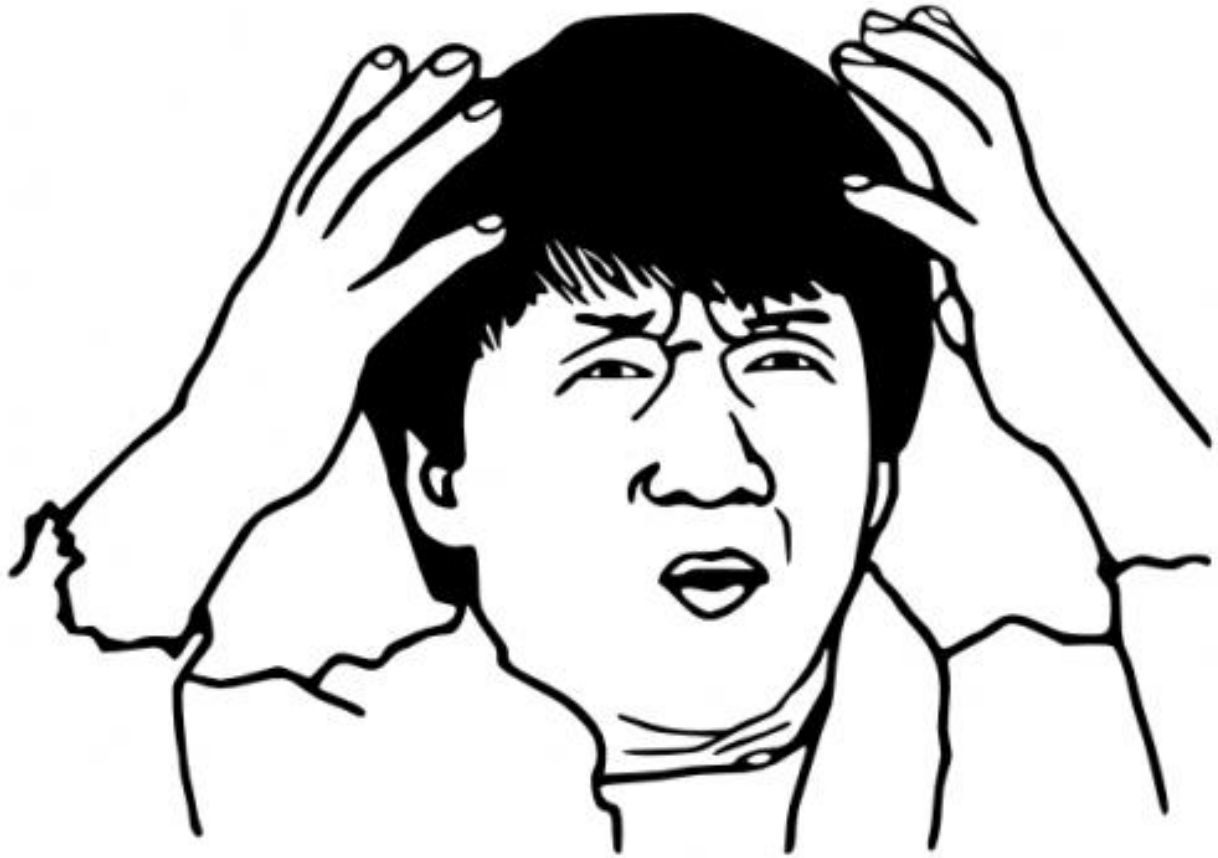
Special thanks to these AAPOR presenters for the following slides: Robin Kaplan, Jennifer Kelley, Victoria Porterfield, Jerry Timbrook, and Georg-Christoph Haas

PAPOR Annual Mini-Conference

June 29, 2018

KFF
HENRY J KAISER FAMILY FOUNDATION

Filling the need for trusted information on national health issues.



KFF
HENRY J KAISER
FAMILY FOUNDATION

Presentation Outline

1. How Do We Define Survey Burden?
2. How Do We Measure Burden and Identify its Correlates?
3. How Can We Make Efforts to Reduce Burden?

KFF
HENRY J KAISER FAMILY FOUNDATION

But Why?

RUTGERS

Edward J. Bloustein School
of Planning and Public Policy

A Diagnostic Mechanism for Assessing Respondent Burden: Sensitive Item Nonresponse Bias in Student Surveys

Victoria Porterfield, Ph.D
Marc D. Weiner, J.D., Ph.D
Paul Siracusa, M.P.P

Edward J. Bloustein School of Planning and Public Policy
Rutgers-New Brunswick

Context: *Institutional Review Boards* [IRB]

- National Research Act of 1974:
  - Every federally funded university in the United States and conducts research is required to have an institutional review board (IRB);
  - Regulations put in place as a response to various misuses of humans in experiments and studies.

- Fast forward to present day:
  - Survey research is considered human subjects research;
  - Sometimes what is considered 'sensitive' is vague and outdated, but often times it is not;
  - How do IRBs determine what is sensitive?
    - This study provides insight into what might (and might not) be considered sensitive.

Results

Probability of Sensitive Item Response

	Income	Religion	Sexual Orientation	Political Affiliation
International	.9074	.9722	.8241	.9583
Domestic	.9465	.9887	.9413	.9773
Asian International	.9128	.9767	.8081	.9593
Domestic and NonAsian International	.9461	.9885	.9409	.9772

- Mean INRR for sexual orientation was 5.4%; International was 14.0%.
- Mean INRR for income was 4.8%; International was 7.4%.

How Do We Define Survey Burden?

KFF
HENRY J KAISER
FAMILY FOUNDATION

Types of Survey Burden

1. Respondent-related
    - Objective measures
    - Subjective measures

2. Interviewer-related
    - Deviations from question wording

KFF
HENRY J KAISER FAMILY FOUNDATION

Survey features and respondent characteristics that contribute to objective and subjective measures of burden

Robin Kaplan and Scott Fricker
Bureau of Labor Statistics, Office of Survey Methods Research
AAPOR 2018

**Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the author(s) and not necessarily those of the BLS*

Objective Measures of Respondent Burden

- Time to complete survey:
  - Survey length
  - Number of questions
  - Frequency of survey
  - Time reading instructions
  - Time gathering and entering data
  - Time reviewing data

Subjective Measures of Respondent Burden

- Appraisal of how burdensome the survey experience was, for example:
  - Effort
  - Motivation
  - Interest
  - Sensitivity

- "How burdensome did you find this survey?"

Fricker et al., 2014

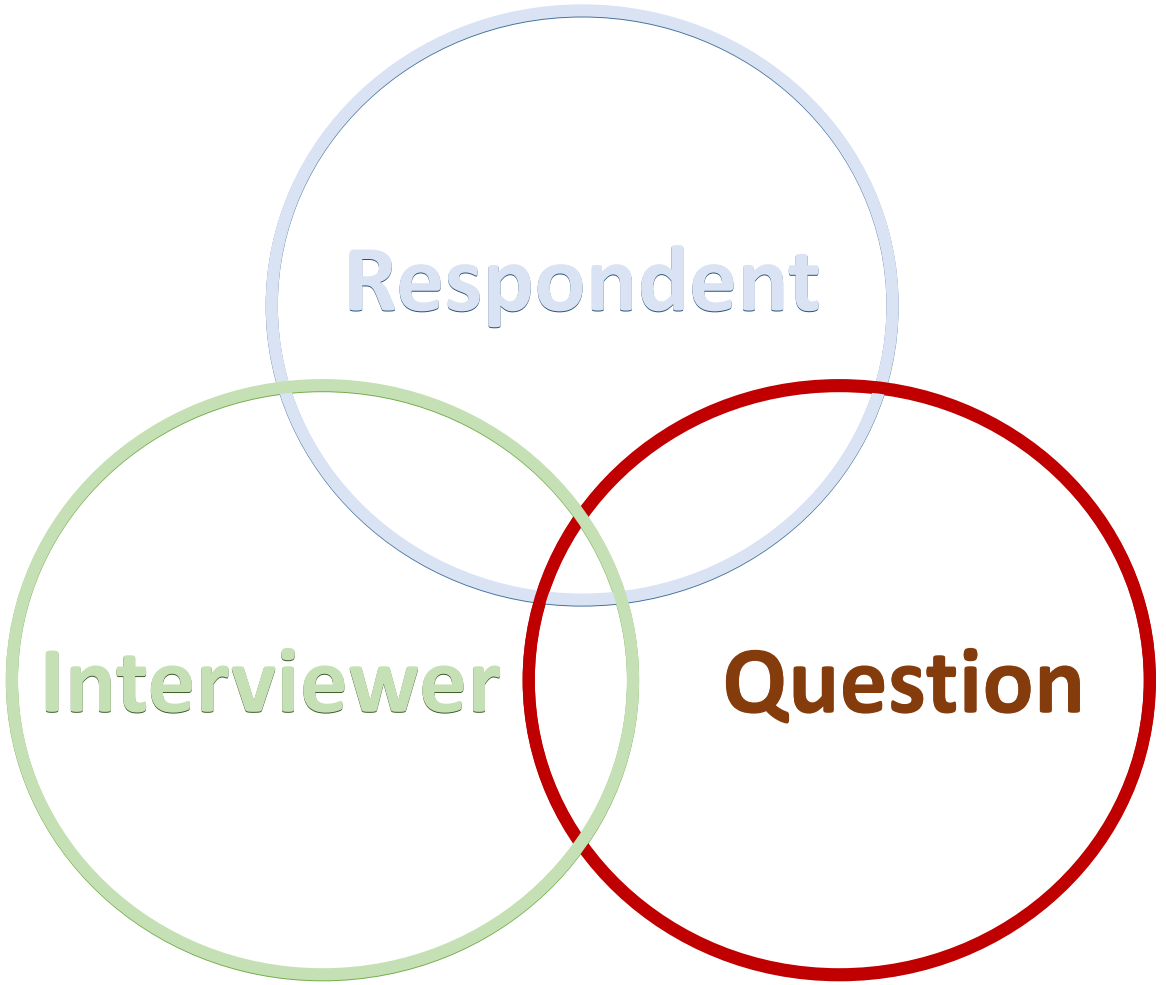
Question Characteristics and Interviewer Question-Reading Deviations

Jennifer Kelley
University of Essex

AAPOR 73rd Annual Conference
May 16th, 2018

Why do interviewers go off script?

- Simple error
- Tailor question to respondent to signal they are listening
- Dealing with uncooperative/tired respondent
- Lack of training
- Personal gain (e.g., paid by the interview)
- Trying to “help” respondent
- “Fix” the question



How Do We Measure Burden and Identify its Correlates?

Survey features and respondent characteristics that contribute to objective and subjective measures of burden

Robin Kaplan and Scott Fricker
Bureau of Labor Statistics, Office of Survey Methods Research
AAPOR 2018

**Disclaimer: This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the author(s) and not necessarily those of the BLS*

Research Questions

- What survey/respondent characteristics contribute to objective & subjective burden? How does burden affect data quality?
  - Does respondents' level of engagement / survey fatigue affect burden? (e.g., McCalin et al., 2015)
  - Does the order of the subjective burden questions affect respondent ratings of the survey? (e.g., Schwarz et al., 1991)

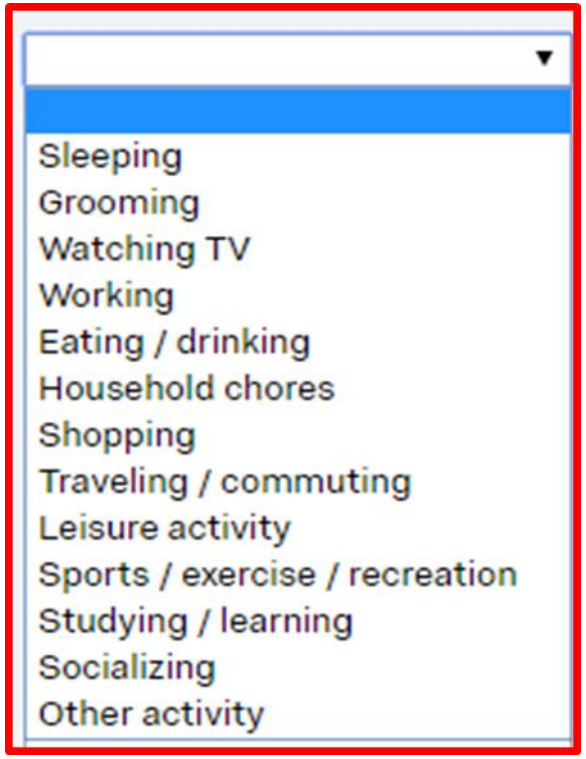
Study Design (single web survey ~20 min)

1. Activity log task
2. Survey questions about typical time use
3. Level of engagement questions
4. Subjective burden ratings
5. Demographic questions

Objective Burden Measures

Average time on survey

Average time on activity log



Subjective Burden Measures (1-5 scales)

Overall burden

How burdensome was it to complete this survey?

Activity Log burden

How burdensome was it to complete the activity log?

Effort

How effortful was it to complete this survey?

Easy/Difficult

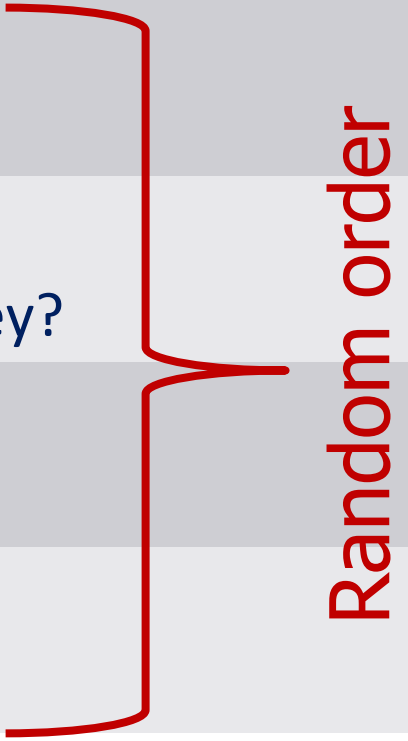
How easy or difficult was it to answer the questions in this survey?

Sensitivity

How sensitive were the questions in this survey?

Interest

How interesting did you find this survey?



Other Subjective Measures

Fatigue	**How well-rested do you feel right now?** 1. Not at all rested 2. A little rested 3. Somewhat rested 4. Very rested
Perception of length	**Did you feel the length of this survey was too short, about right, or too long?** 1. Too short 2. About right 3. Too long

Engagement measure

Please indicate to what degree you were having each experience described below **while you completed the survey**. Please answer according to what really reflected your experience rather than what you think your experience should have been. [1 *strongly disagree* to 5 *strongly agree*]

1.) I was finding it difficult to stay focused on the survey.

2.) I was doing the survey without paying attention.

3.) I was preoccupied with the future or the past.

4.) I was doing the survey automatically, without being aware of what I was doing.

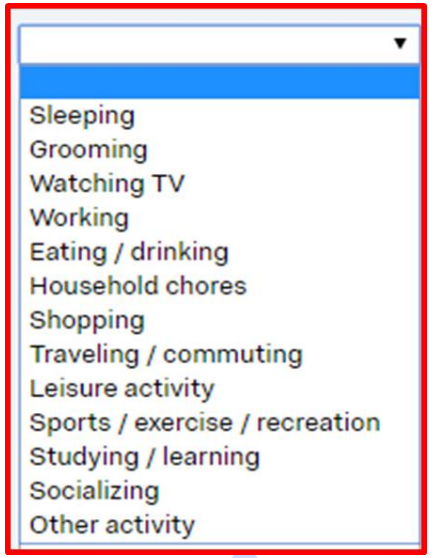
5.) I was rushing through the survey without really being attentive to it.

Brown & Ryan (2003)

Objective Burden Outcomes

Average time on survey*	20.89 min (SD = 22.70)
Average time on activity log	6.69 min (SD = 5.13)

*Removed n=17 outliers
(participants who took more than 3 standard deviations above the average time to complete the survey)



Subjective Burden Measures (1-5 scales)	Mean	SD
Overall burden How burdensome was it to complete this survey?	1.36	0.66
Activity Log burden How burdensome was it to complete the activity log?	1.79	0.89
Effort How effortful was it to complete this survey?	2.29	1.01
Easy/Difficult How easy or difficult was it to answer the questions in this survey?	1.84	0.83
Sensitivity How sensitive were the questions in this survey?	1.53	0.87
Interest How interesting did you find this survey?	3.11	1.10

Other Subjective Measures

		Mean	SD
Fatigue	**How well-rested do you feel right now?** 1. Not at all rested 2. A little rested 3. Somewhat rested 4. Very rested	2.86	0.94
Perception of length	**Did you feel the length of this survey was too short, about right, or too long?** 1. Too short 2. About right 3. Too long	2.12	0.33

Engagement measure (1 -5 scale)

*All items recoded. Higher scores = more engagement

Overall Mean = 4.44 (SD = 0.76; α = 0.90)	Mean	SD
I was finding it difficult to stay focused on the survey.	4.21	1.04
I was doing the survey without paying attention.	4.57	0.81
I was preoccupied with the future or the past.	4.27	1.02
I was doing the survey automatically, without being aware of what I was doing.	4.54	0.84
I was rushing through the survey without really being attentive to it.	4.60	0.79

Regression Models

Step 1 - Objective Burden Measures
Step 2 - Subjective Burden Measures
Step 3 - Survey Engagement
Step 4 - Respondent Demographics

<u>Outcomes:</u>
Objective burden, Subjective burden, Data quality

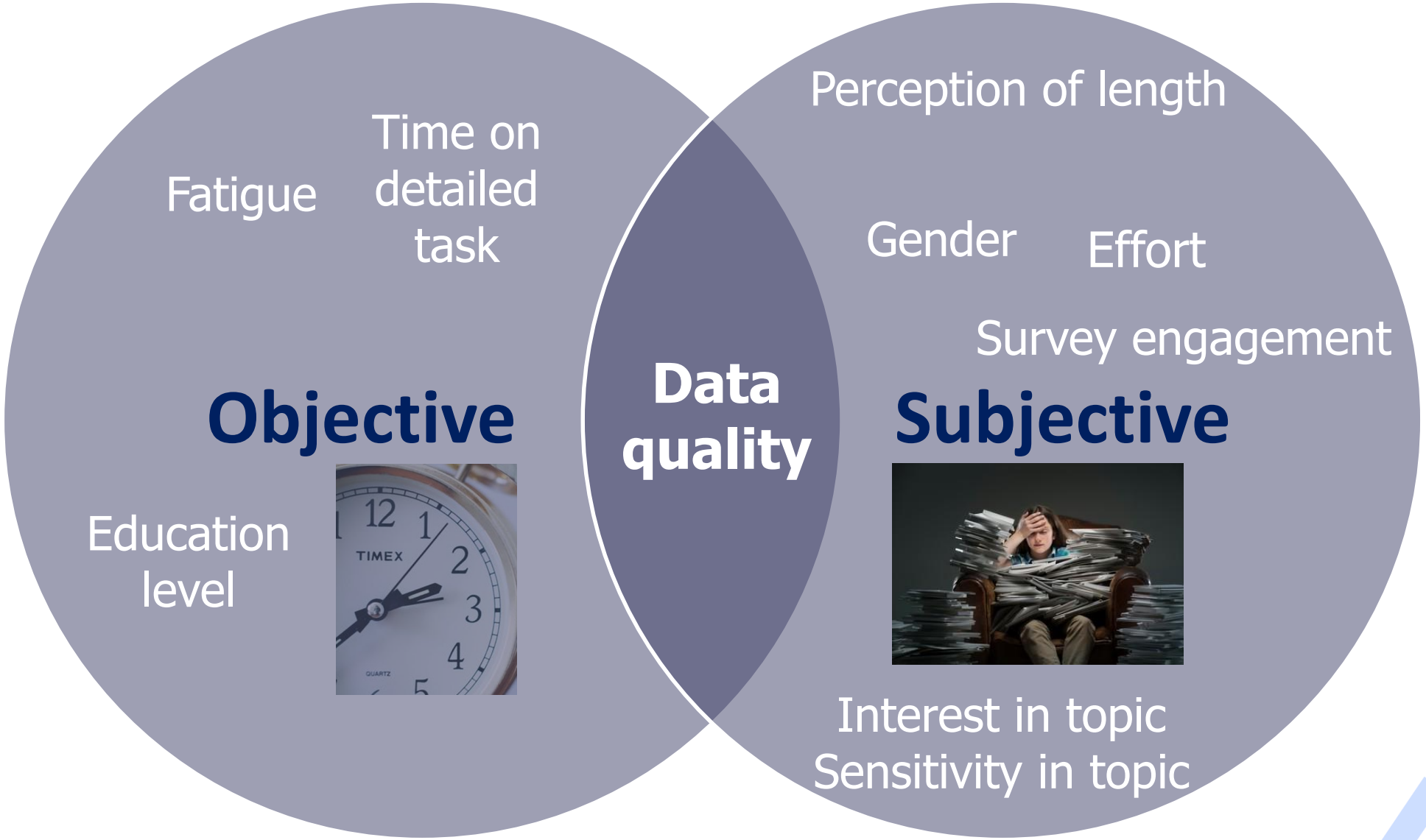
Subjective Burden ("How burdensome was it to complete this survey?")

Predictor	*B*	p-value
Time Spent on Survey	0.06	0.085
Time Spent on Activity Log (mean centered)	0.01	0.936
Burden-Activity Log**	**0.50**	**< .001**
Effort*	**0.08**	**0.002**
Easy/Difficult	0.04	0.276
Interest*	**-0.07**	**0.029**
Sensitivity*	**0.10**	**0.001**
Well-Rested	-0.01	0.577
Perception of Survey Length**	**0.10**	**< .001**
Survey Engagement**	**-0.12**	**< .001**
Gender (0 = male; 1 = female)	-0.05	0.054
Age	0.01	0.640
Education Level (0 = no degree; 1 = Associate's or higher degree)	0.04	0.134

** p < .001
*p< .05

Note: Adjusted R^2=.47; $F(18, 985) = 48.76$, $p < .001$ at Step 4.

Summary


Fatigue
Time on detailed task
Objective
Education level
Data quality
Perception of length
Gender
Effort
Survey engagement
Subjective
Interest in topic
Sensitivity in topic

Conclusions & Future Directions

- Conclusions
  - More than just time contributed to data quality outcomes
  - Easing respondent burden
    - Survey engagement/interest
- In the future:
  - Burden as a complex, multi-component concept
  - Continue to assess objective and subjective burden within surveys to better understand burden and its relationship to data quality, survey outcomes

Question Characteristics and Interviewer Question-Reading Deviations

Jennifer Kelley
University of Essex

AAPOR 73rd Annual Conference
May 16th, 2018

Research Question

Are there certain question characteristics that increase the odds of interviewers engaging in *major* question-reading deviations in *face-to-face* interviews?

Data

- Wave 3 of the Understanding Society Innovation Panel, U.K.
  - Multi-stage probability sample
  - Social and economic topics
- 1621 CAPI interviews
- Interviewers are trained to read all questions verbatim
- Sections of the interview were recorded with permission of respondent
  - 1167 respondents gave consent
- Interview recordings
  - 820 recordings were available for analysis
- Interviewers were told which sections would be recorded

Methods: Sample

- Randomly selected two recorded interviews from each interviewer (n=80) 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 same response options for all regions
- If the two interviews selected for the interviewer did not result in a minimum of 50 questions, a third interview was selected
- 10,345 questions

Methods: Behavior Coding for Question-Reading

- Interviewer's first reading of the question was coded
- Verbatim or Deviation
- Magnitude of deviation
  - Minor – does not change question meaning
  - Major – changes question meaning

Question Reading (n=10345)	%
Verbatim	52.5
Minor Deviation	34.5
Major Deviation	13.0

Methods: Coded Question Characteristics

- Word Count
- Difficulty (FRQ)
- Order of Question
- Type of Question
- Confirming Past Information
- Double-barreled
- Sensitive
- Type of Response Options
- Response Options Read in Text
- Number of Response Options
- Gate or Follow up Question
- Common Stem
- Part of Series
- Place in Series
- Optional Text
- Definition or Example
- Time Reference
- Showcard
- Scripted Help Text

Analysis Methods

- Bivariate analysis to assess the relationship between the question characteristics and major deviations

- Logistic multi-level model
  - Three levels: where *n* questions (Level 1) are nested within respondents (Level 2) nested within interviewers (Level 3)
  - Outcome variable: Major question-reading deviation
  - Predictor variables: 19 Question Characteristics
  - Control variables: Respondent and Interviewer level variables (*next slide*)
  - Estimated using RIGLS (PQL-2) in MLwiN 3.01
  - Presenting odds ratios

Methods: Control Variables

- Respondent Characteristics
  - Age
  - Education
  - Married
  - Employed
  - Number of Children in Home
  - Nationality (British/Other)
  - Cognitive Ability
  - Completed Interview Last Wave
- Interview Context
  - Interviewer’s Assessment of R's Understanding
  - Interviewer’s Assessment of Resistance
  - Other Present
  - Number of Interviews Same Day
- Interviewer Characteristics
  - Age
  - Nationality (British/Other)
  - Experience
  - Average Number of Interviews per Day

Odds Ratios of Interviewer Making Major Question-Reading Deviations

Order in Questionnaire	1.003****
Word Count	1.007**
Difficulty (FKG)	1.064****
Type of Question (ref=Intro/Instruct)	
Attitude	0.689
Behavioral	1.728*
Demo/Fact	2.477****
Optional Text	0.168****
Gate Questions (ref=Independent Question)	
Gate	1.404**
Follow-up	1.828****
Place in Series	0.894*
Response Options Read in Question	5.083****
Type of Response (ref=Other)	
Y/N	1.103
Select 1	0.339****
Select all	0.279****
Scale	0.354*
Definition/Example	7.683****
Time Reference	1.619***
Showcard	1.768*
Question Help	1.582****

*p<0.05
**p<0.01
***p<0.001
****p<0.0001

Discussion

- Questions with definitions/examples and questions with response options read as part of the text have the highest odds of question-reading deviations
- Question aids may be doing more harm than good
- Questionnaire designers need to be aware of the risks
  - May want to redesign question
  - When training interviewers, draw attention to these types of questions and convey importance of reading the questions verbatim

- Next steps
  - Look at interactions
  - Other respondent or interviewer characteristics

The Effects of Question Complexity and Necessary Question Features on Interviewer and Respondent Contributions to Response Time

Jerry Timbrook
Kristen Olson
Jolene Smyth
University of Nebraska-Lincoln
AAPOR, May 2018

Response Times

- **The problem:**
  - Active timers aggregate respondent and interviewer contributions to response time
    - Cannot investigate respondent and interviewer contributions to response time separately
  - Factors affecting response time (e.g., question characteristics) may affect respondents and interviewers differently
    - Actors have different jobs and different cognitive processes

Factors Affecting Response Time

- **In this study, we:**
  - Use a new timer method
    - Audio record and code respondents' and interviewers' contribution to response time separately
  - Explore the effect of:
    1. Respondents' answering device
    2. Essential question features
    3. Question complexity
  - Compare new method to the previous "total time of both actors" approach
    - Assess the benefits of decomposing response time by actor

Data

- **US/Japan Newspaper Opinion Poll**
  - National telephone survey of U.S. adults conducted by Gallup in Nov 2013.
  - Landline and Mobile, AAPOR RR1 = 7.4%
  - Stratified random subset of 434 interviews were recorded and transcribed
    - 249 landline respondents, 185 mobile respondents

Data

- **DVs:**
  - Total Time:
    - Both Actors Spent on Each Question
    - Respondents Spent on Each Question
    - Interviewers Spent on Each Question
  - Calculated using Sequence Viewer (Dijkstra 1999)
  - 73 items
    - 29,514 Total times (seconds)
    - 29,514 Respondent times (seconds)
    - 29,514 Interviewer times (seconds)

Data

- **IV: Respondent's survey completion device**
  - Coded from respondent self-reports (=0 landline; =1 cell phone)

- **IV: Necessary Question Features**
  - Number of words in the stem and response options
  - Interviewer instructed to read response options? (=0 no; =1 yes)
  - Question type (=0 Attitude/Opinion; =1 Factual)

Data

- **IV: Question Complexity**
  - Reading Level (Flesch–Kincaid Grade Level)
  - Item-Level indicators of QUAID-identified problems:
    - Uncommon Technical Terms (n=42)
    - Vague Words (n=36)
    - Complex Syntax (n=3)
    - Working Memory Overload (n=11)
- **Controls:**
  - Respondent age and education, interviewer tenure

Factors Affecting Response Time

- **Respondents' Answering Device**
  - Cell phone interviews last longer than landline interviews (Kennedy & Everett 2011; Lynn & Kaminska 2012; Timbrook et al. forthcoming)
  - Issues with line quality (Timbrook et al. forthcoming)
    - Affects actors' ability to hear and understand one another
  - Only investigated at survey level
  - Hypotheses:

Respondent Total Time	Interviewer Total Time
Cell Phone > Landline	Cell Phone > Landline

Factors Affecting Response Time

- **Necessary Question Features**
  - Longer questions take longer to administer (Yan and Tourangeau 2008; Couper and Kreuter 2013; Olson and Smyth 2015)
    - More words and more response options
    - Interviewer reading response options?
  - Question Type:
    - Attitude questions take longer than factual questions (e.g., demographics)
  - Hypotheses:

Respondent Total Time	Interviewer Total Time
Longer Questions > Shorter Questions	Longer Questions > Shorter Questions
Read RO = Don't Read RO	Read RO > Don't Read RO
Attitude > Factual	Attitude = Factual

Factors Affecting Response Time

- **Question Complexity**
  - More complex questions have longer response times in:
    - Web (Yan & Tourangeau 2008)
    - Face-to-face (Couper & Kreuter 2013)
    - Landline CATI surveys (Olson & Smyth 2015)

Factors Affecting Response Time

- **Question Complexity**
  - Question Understanding Aid (QUAID) (Graesser et al. 2006)
    - Identifies problems with a question that may affect respondent comprehension
    - Measures:
      - Uncommon technical terms
      - Terms with vague meanings (e.g., *many*, *few*)
      - Complex syntax
      - Working memory overload

Factors Affecting Response Time

- **Question Understanding Aid (QUAID)** (Graesser et al. 2006)
  - How might questions with QUAID-identified problems affect cognitive processing?
    - Respondents
      - Higher working memory burden in telephone surveys (de Leeuw 2005; Dillman et al. 2014)
    - Interviewers
      - Question asking (Japec 2008; Dykema et al 2017)
      - Feedback (Japec 2008)
- **Hypotheses:**

Respondent Total Time	Interviewer Total Time
QUAID-identified Problem > No Problem	QUAID-identified Problem > No Problem

Factors Affecting Response Time

- **Question Complexity**
  - Question reading level
    - Higher reading levels (harder to understand) = longer response times (Yan & Tourangeau 2007; Olson & Smyth 2015)
  - Hypotheses:

Respondent Total Time	Interviewer Total Time
↑ Reading Level > ↓ Reading Level	↑ Reading Level > ↓ Reading Level

Modeling Approach

- **Cross-classified random effects models**

$$log\left(Y_{i(j_1,j_2)k}\right) = \gamma_{0000} + v_{000k} + u_{0j_10k} + u_{00j_2k} + \varepsilon_{i(j_1,j_2)k}$$

- **Estimate models predicting:**
  - log(Total Time on a Question)
  - log(Respondent Time on a Question)
  - log(Interviewer Time on a Question)

Results

	Variable	Total Time on Qn	Resp. Time on Qn	I'wer Time on Qn
Device	Cell Phone (ref=Landline)	+***	+***	+***
Necessary Question Features	# of Words in Stem	+***	n.s.	+**
	# of Words in RO	n.s.	n.s.	n.s.
	Read RO (ref=Did not)	+***	n.s.	+***
	Factual Question (ref=Attitude)	n.s.	n.s.	n.s.
Question Complexity	Reading Level	+*	n.s.	+*
	Technical Terms	n.s.	n.s.	+**
	Vague Words	+***	n.s.	+***
	Complex Syntax	n.s.	n.s.	n.s.
	Working Memory Overload	n.s.	n.s.	+*

*p<.05, **p<.01, ***p<.001

How Can We Make Efforts to Reduce Burden?

KFF
HENRY J KAISER
FAMILY FOUNDATION

Institute for Employment Research
The Research Institute of the Federal Employment Agency

IAB

Assessing the Effect of the Web Mode on Response Burden in Establishment Surveys

AAPOR, May 17th 2018

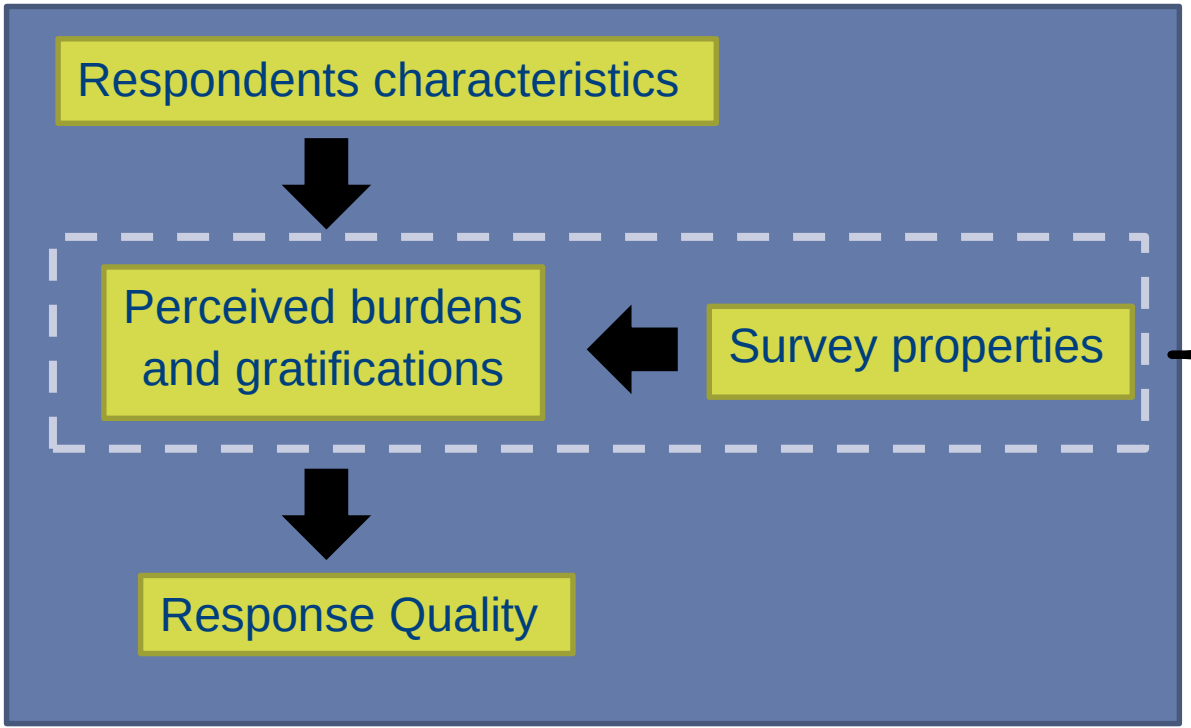
Georg-Christoph Haas
(IAB Nuremberg, Germany)

Stephanie Eckman
(RTI International)

Ruben Bach
(University of Mannheim, Germany)

Frauke Kreuter
(IAB Nuremberg, Germany)

Response Burden Model (Haraldsen 2004)



Web vs mail mode

- **Mode**
- Content
- Questionnaire length
- Survey Organisation
- Publicity
- Implementation
- Strategy
- Content
- Layout
- Communication strategy
- Sample design

The Effect of Web on Response burden

Increase Burden

- Poor online skills
- Poor web survey design
- Plausibilty checks

Decrease Burden

- Invisible filters
- Counting and calculating
- Paperwork burden
- Access information more quickly

Data

- Sample drawn from German admin records (16,000)
  - Stratified by **size, location and industry class**
- 3 different mode conditions
- 1,574 completes (AAPOR RR1: 10.2 %)

Experimental Design

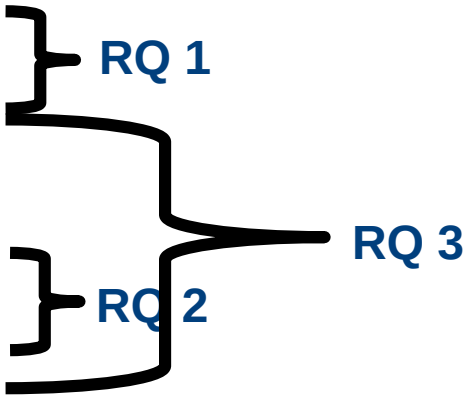
- Cover letter with link and password
- Paper questionnaire
- Return envelope

Mode	1. Mailing	2		Initial sample size	RR in %
Mail only	Mail	Mail	Mail	4,000	11.7
Web only	Web	Web	Web	4,000	5.6
Choice	Mail/Web	Mail/Web	Mail/Web	8,000	11.8

All invitation letters were sent by mail and included a data protection sheet

Research Questions

Mode	N
Mail	492
Web	241
Mail of Choice	650
Web of Choice	321



RQ 1: **Mail vs Web**

RQ 2: **Mail vs Web**, Choice impacts response burden

RQ 3: **Web vs Web of Choice**, Choice impacts response burden

Dependent Variables (Dale et al 2007)

- **Estimated Time** (in minutes)
  - Time to gather needed information
  - Time to complete the questionnaire

- **Perceived Burden** (5 Item scales)
  - Perceived time (very quick – very time consuming)
  - Perceived burden (very easy – very burdensome)

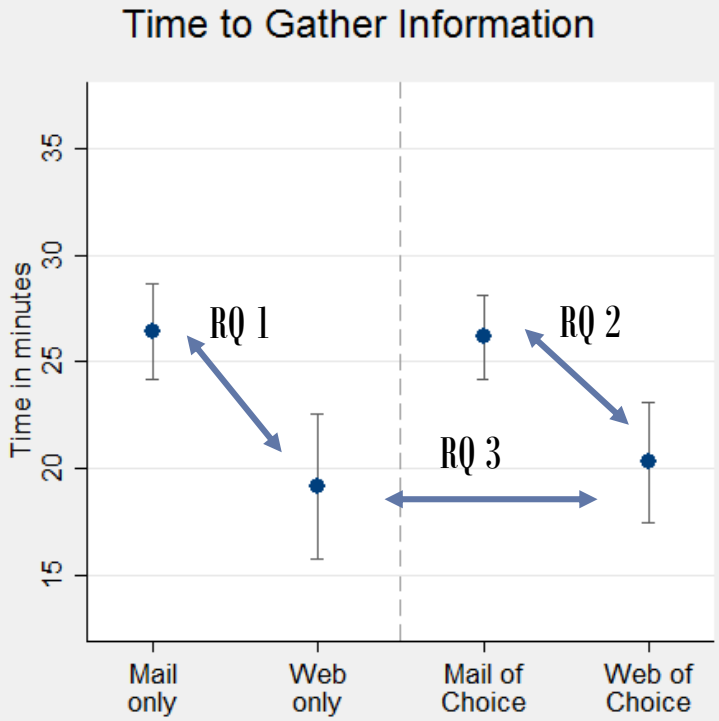
Models

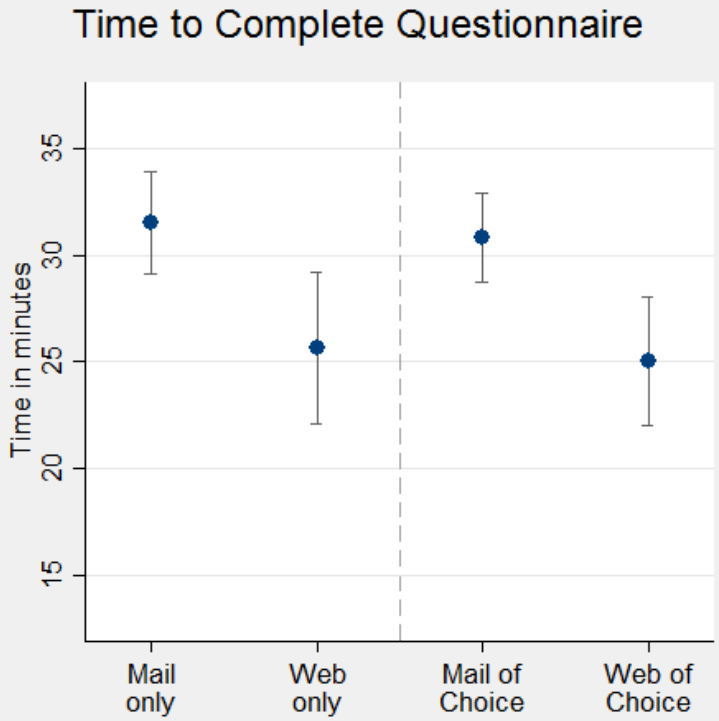
	Model	DV	IV	Controls

Estimated Time	Median Regression	• Time to gather
		• Time to complete
			• Mail
			• Web
			• Mail of Choice
			• Web of Choice
				• Questionnaire length
				• Content
				• Establishment
				  • Size
				  • Industry
				  • Location

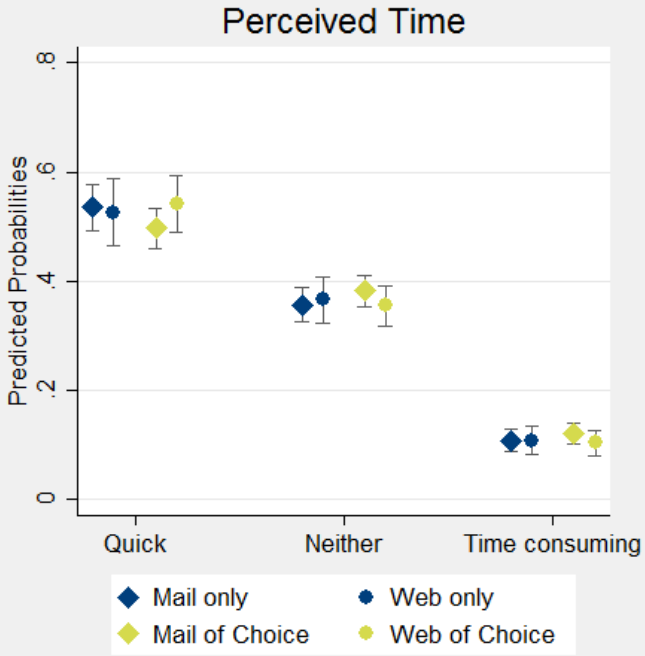
Perceived	Ordinal logistic regression	• Perceived time • Perceived burden		

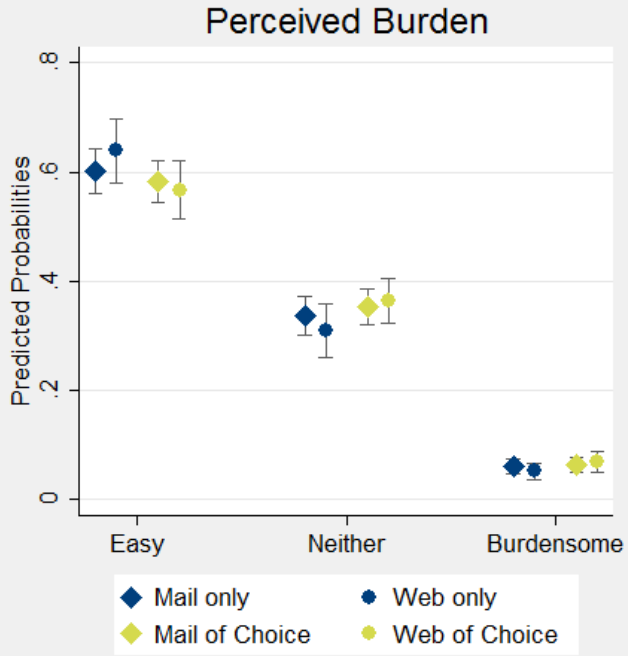
Results – Estimated Time





Results – Perceived Burden Indicators





The Effects of Question Complexity and Necessary Question Features on Interviewer and Respondent Contributions to Response Time

Jerry Timbrook
Kristen Olson
Jolene Smyth
University of Nebraska-Lincoln
AAPOR, May 2018

Factors Affecting Response Time

- **Question Complexity**
  - Question Understanding Aid (QUAID) (Graesser et al. 2006)
    - Identifies problems with a question that may affect respondent comprehension
    - Measures:
      - Uncommon technical terms
      - Terms with vague meanings (e.g., *many*, *few*)
      - Complex syntax
      - Working memory overload

Conclusions – Question Complexity

Respondent Total Time	Interviewer Total Time
QUAID-identified Problem > No Problem ✓	QUAID-identified Problem > No Problem
↑ Reading Level > ↓ Reading Level ✓	↑ Reading Level > ↓ Reading Level

- **For respondents, questions with QUAID-identified problems and reading level had no effect on response time**
  - Some respondent processing time absorbed by interviewer speaking time?
- **Interviewers take longer on questions with QUAID-identified problems and higher Reading Levels**
  - Interviewers are doing the "spoken work" for these problematic questions
  - Use the QUAID tool and reading level to evaluate and revise questions to reduce burden

Concluding Remarks

1. There are a number of ways to define and characterize survey burden
   - We tend to think of it in terms of respondents' experiences when completing a survey, but burden can also be associated with the interviewer

2. Many methods to measure burden as well as identify the features of the survey and characteristics of respondents and interviewers that are correlated with it
   - For example, reading response options ↑ reading deviations and reading time

3. Some possible ways to reduce burden
   - Implementing QUAID and testing out different survey modes

Thank you.