What do interviewers learn?
An examination of interview length and interviewer behaviors

Kristen Olson and Jolene Smyth
University of Nebraska-Lincoln
Interviewers and their Effects in a Total Survey Error Framework Workshop
Lincoln, NE
February 2019
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 Antje Kirchner, Beth Cochran, Jinyoung Lee, Amanda Ganshert, and Jerry Timbrook for research assistance!

• Thanks to all of our transcriptionists and behavior coders for their amazing work!
Interviewers set the stage for respondents

• Interviewers are important actors in telephone surveys
• By setting the pace for an interview, interviewers communicate the amount of time and cognitive effort respondents should put into their task
• But interviewers vary widely in the time they spend administering a survey
  • And it changes over the course of the data collection period as the interviewer gains within-study experience (e.g., Olson and Peytchev 2007; Olson and Bilgen 2011)
    • In particular, they speed up.
• We don’t know what leads to these differences in speed of administering a questionnaire.
Three hypotheses

• Certain interviewer behaviors are omitted or shortened over the course of a field period.
  • Standardized “good” behaviors go away (Ongena and Dijkstra 2007; Olson and Bilgen 2011; Tarnai and Moore 2008; van der Zouwen, Dijkstra and Smith 1991)
• Interviewers may not change the prevalence of individual “good” behaviors, but become more efficient in them or eliminating extraneous behaviors (Olson and Peytchev 2007; Cleary, Mechanic and Weiss 1981; Houtkoop-Steenstra 1997).
  • Not directly trained, but happens over the course of interviews.
• Increased use of bad behaviors that shortcut time (van der Zouwen et al. 1991)
  • Nonstandardized bad behaviors that always happen.
Kirchner and Olson (2017, JSSAM)

• What explains interview length?
  • interviewer Experience
    • Learning, overall experience, and interviewer cooperation rate
  • Response propensity
    • Composition: Respondent gender, age, education, race, employment status, income, HH size, parent, volunteer status
    • Contactability and cooperation: Item NR rate, ever refusal, complete at first contact, # of call attempts, time of day interview completed
    • Interaction between R and I: Word count of interview

• But there is much more to the interaction between the R and I than just the number of words that they speak
This paper

• What interviewer behaviors change over the course of the data collection period in two telephone surveys?

• Do these behaviors account for changes in survey length over the course of the data collection period?
Data – Building off Kirchner and Olson (2017)

• Work and Leisure Today 1 Survey
  • Landline RDD CATI survey
  • Conducted by AbtSRBI between July 31 and August 28, 2013
  • N=450, AAPOR RR3=6.3%
  • Questionnaire deliberately designed to have highly problematic questions
  • Data deposited at ICPSR; under review

• Work and Leisure Today 2 Survey
  • Dual Frame RDD CATI survey
  • Conducted by AbtSRBI during September 2015
  • n=902, Landline = 451, AAPOR RR3=9.4%; Cell phone = 451, AAPOR RR3=7.1%
  • Two versions – alternative experimental questionnaire designs
  • Questionnaire deliberately avoided these highly problematic questions
**Question text:** How much do you enjoy cooking? Not at all, a little, somewhat, a lot, or completely?

<table>
<thead>
<tr>
<th>Transcripts</th>
<th>Actor</th>
<th>Initial</th>
<th>Assessment</th>
<th>Details</th>
<th># seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: And how much do you enjoy cooking? Not at all, a little, somewhat, a lot, or completely?</td>
<td>Interviewer</td>
<td>iQuestion Asked</td>
<td>Read exact</td>
<td></td>
<td>4.7</td>
</tr>
<tr>
<td>R: Um, how, what? I didn't catch--.</td>
<td>Respondent</td>
<td>rClarification</td>
<td>Asks for repeat of question</td>
<td></td>
<td>4.7</td>
</tr>
<tr>
<td>I: How much do you enjoy cooking?</td>
<td>Interviewer</td>
<td>iProbes</td>
<td>Repeat part of Q exact</td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>R: Cooking? I love to.</td>
<td>Respondent</td>
<td>rAnswer Provided</td>
<td>Uncodable answer</td>
<td>rElaborates no implied</td>
<td>1.4</td>
</tr>
<tr>
<td>I: Okay.</td>
<td>Interviewer</td>
<td>Feedback</td>
<td>Affirmation</td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>R: That's, that's my favorite hobby.</td>
<td>Respondent</td>
<td>Feedback</td>
<td>Personal disclosure</td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>I: Okay, so a lot or completely?</td>
<td>Interviewer</td>
<td>iProbes</td>
<td>Asks for explicit response</td>
<td>Probe directively, no mismatch</td>
<td>1.3</td>
</tr>
<tr>
<td>R: Uh, I'd say a lot. I'm thinking about going to culinary school.</td>
<td>Respondent</td>
<td>rAnswer Provided</td>
<td>Adequate answer</td>
<td>rAdequate w elaboration</td>
<td>3.4</td>
</tr>
<tr>
<td>I: Oh, good for you.</td>
<td>Interviewer</td>
<td>Feedback</td>
<td>Short acknowledgement</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Behavior Codes

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

<table>
<thead>
<tr>
<th></th>
<th>Actor</th>
<th>Initial Action</th>
<th>Assessment of Initial Action</th>
<th>Details of Action</th>
<th>Parentheses</th>
<th>Laughter</th>
<th>Disfluencies</th>
<th>Interruptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>WLT1</td>
<td>κ=0.998</td>
<td>0.90</td>
<td>0.55 to 0.68</td>
<td>0.10 to 0.77</td>
<td>0.92</td>
<td>0.96</td>
<td>0.87</td>
<td>0.94</td>
</tr>
<tr>
<td>WLT2</td>
<td>κ=0.998</td>
<td>0.93</td>
<td>0.36 to 0.76</td>
<td>0.24 to 0.83</td>
<td>0.95</td>
<td>0.97</td>
<td>0.83</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Creating behavior measures

• Two ways of examining measures of behaviors
  • Conversational turn level – Total number of conversational turns on which a behavior occurred
    • This is a measure of how much conversation occurred due to this behavior
    • Some questions can have multiple turns with the same kind of behavior (e.g., multiple probing turns)
  • Question level – Total number of questions on which a behavior occurred
    • This is a measure of how spread out across the questionnaire each behavior was

• Obviously highly correlated
  • Focus on questions in this presentation. Results are similar for conversational turns.
How much do you enjoy cooking? Not at all, a little, somewhat, a lot, or completely?

<table>
<thead>
<tr>
<th>Initial</th>
<th>Assessment</th>
<th>Details</th>
<th># seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>iQuestion Asked</td>
<td>Read exact</td>
<td></td>
<td>4.7</td>
</tr>
<tr>
<td>rClarification</td>
<td>Asks for repeat of question</td>
<td></td>
<td>4.7</td>
</tr>
<tr>
<td>iProbes</td>
<td>Repeat part of Q exact</td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>rAnswer Provided</td>
<td>Uncodable answer</td>
<td>rElaborates no implied</td>
<td>1.4</td>
</tr>
<tr>
<td>Feedback</td>
<td>Affirmation</td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>Feedback</td>
<td>Personal disclosure</td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>iProbes</td>
<td>Asks for explicit response</td>
<td>Probe directly, no mismatch</td>
<td>1.3</td>
</tr>
<tr>
<td>rAnswer Provided</td>
<td>Adequate answer</td>
<td>rAdequate w elaboration</td>
<td>3.4</td>
</tr>
<tr>
<td>Feedback</td>
<td>Short acknowledgement</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Dependent Variables

- Interview length in minutes, trimmed at 1<sup>st</sup> and 99<sup>th</sup> percentiles
  - WLT1: 12.65 minutes
  - WLT2: 13.36 minutes

- Interviewer behaviors
  - Standardized “good” behaviors
    - Exact question reading; Nondirective probes; Exact verification; Appropriate clarification; Appropriate feedback
  - Efficiency behaviors
    - Stuttering during question reading; Disfluencies; Pleasant talk; Task-related feedback; Laughter
  - Nonstandardized “bad” behaviors
    - Minor changes in question wording; Major changes in question wording; Directive probes; Inadequate verification (paraphrasing); Interruptions
Standardized behaviors

- Exact Question Reading: WLT1 23.4, WLT2 37.0
- Nondirective Probes: WLT1 8.4, WLT2 7.7
- Exact Verification: WLT1 7.9, WLT2 7.2
- Appropriate Clarification: WLT1 1.9, WLT2 0.4
- Appropriate Feedback: WLT1 19.5, WLT2 22.6
Efficiency behaviors

- Stuttering during q'n reading: WLT1 = 2.7, WLT2 = 2.4
- Disfluencies: WLT1 = 13.3, WLT2 = 11.7
- Pleasant talk: WLT1 = 0.5, WLT2 = 0.7
- Task-related feedback: WLT1 = 0.9, WLT2 = 1.4
- Laughter: WLT1 = 2.3, WLT2 = 3.1
Nonstandardized behaviors

![Bar chart showing the number of questions with different behaviors for WLT1 and WLT2.]

- Any changes in question wording: WLT1 = 20.7, WLT2 = 11.1
- Minor changes in question wording: WLT1 = 15.4, WLT2 = 4.9
- Major changes in question wording: WLT1 = 5.4, WLT2 = 6.3
- Directive probes: WLT1 = 2.6, WLT2 = 1.1
- Inadequate verification: WLT1 = 3.2, WLT2 = 1.8
- Interruptions: WLT1 = 5.6, WLT2 = 3.3
Primary Independent variable: Within-survey experience

• Log-transformed ordinal counter for within-survey experience
  • WLT1: Ranges from 1 to 27
  • WLT2: Ranges from 1 to 79

• Control variables
  • Overall interviewer experience
  • Interviewer-level cooperation rate, item NR rate, whether R ever refused, complete at first contact, # call attempts, time of day I’w completed
  • Number of questions asked
  • Number of answer changes
  • Respondent sex, age, education employment status, income HH size, parental status, volunteer status, computer usage
  • Interviewer race, gender, interviewer worked primarily weekday evening shifts
  • Version indicator and cell phone interview indicator for WLT2
Analytic strategy

- Two-level random intercept models
  - Poisson models for the interviewer behaviors
    - Number of questions as the exposure variable
  - Linear models for interview length

\[
\text{Log}(IwBehaviors)_{ij} = \gamma_{00} + \beta_1 \text{Ln}(IwOrder)_{ij} + \beta_p \text{Controls}_{pij} + u_{0j}
\]

\[
\text{Length}_{ij} = \gamma_{00} + \beta_1 \text{Ln}(IwOrder)_{ij} + \beta_2 IwBehaviors_{ij} + \beta_p \text{Controls}_{pij} + u_{0j} + \epsilon_{ij}
\]

- Estimated using Stata 15.1 `mepoisson` and `mixed`
This paper

• What interviewer behaviors change over the course of the data collection period in two telephone surveys?

• Do these behaviors account for changes in survey length over the course of the data collection period?
Predicting behaviors in each study

• Focus only on interview order (within-survey experience) coefficient

\[
\log(\text{IwBehaviors})_{ij} = \gamma_{00} + \beta_1 \log(\text{IwOrder})_{ij} + \beta_p \text{Controls}_{p_{ij}} + u_{0j}
\]
Standardized Interviewing Behavior: Associated with within-survey experience?

<table>
<thead>
<tr>
<th></th>
<th>WLT1</th>
<th>WLT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact question reading</td>
<td>0.017</td>
<td>0.001</td>
</tr>
<tr>
<td>Nondirective probes</td>
<td>0.020</td>
<td>-0.033*</td>
</tr>
<tr>
<td>Exact verification</td>
<td>0.020</td>
<td>-0.051**</td>
</tr>
<tr>
<td>Appropriate clarification</td>
<td>0.091+</td>
<td>-0.034</td>
</tr>
<tr>
<td>Appropriate feedback</td>
<td>-0.010</td>
<td>-0.035****</td>
</tr>
</tbody>
</table>

n/s = not significant; +p<.10, * p<.05, ** p<.01, *** p<.001, **** p<.0001
Decreases in standardized behaviors as interviewers gain within-study experience in WLT2; No change in WLT1.
Efficiency Behaviors: Associated with within-study experience?

<table>
<thead>
<tr>
<th></th>
<th>WLT1</th>
<th>WLT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stuttering during q’n reading</td>
<td>-0.201****</td>
<td>-0.201****</td>
</tr>
<tr>
<td>Disfluencies</td>
<td>-0.062**</td>
<td>-0.058****</td>
</tr>
<tr>
<td>Pleasant talk</td>
<td>-0.086</td>
<td>-0.098</td>
</tr>
<tr>
<td>Task-related feedback</td>
<td>-0.151*</td>
<td>-0.052</td>
</tr>
<tr>
<td>Laughter</td>
<td>-0.162****</td>
<td>-0.084****</td>
</tr>
</tbody>
</table>

n/s = not significant; * p<.05, ** p<.01, *** p<.001, **** p<.0001
Fewer efficiency behaviors as interviewers gain within-study experience
Nonstandardized Behaviors: Associated with within-study experience?

<table>
<thead>
<tr>
<th>Behavior</th>
<th>WLT1</th>
<th>WLT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any changes in question wording</td>
<td>0.009</td>
<td>0.043**</td>
</tr>
<tr>
<td>Minor changes in question wording</td>
<td>0.032+</td>
<td>0.050*</td>
</tr>
<tr>
<td>Major changes in question wording</td>
<td>-0.050+</td>
<td>0.036*</td>
</tr>
<tr>
<td>Directive probes</td>
<td>0.100*</td>
<td>-0.087*</td>
</tr>
<tr>
<td>Inadequate verification</td>
<td>-0.176****</td>
<td>-0.106****</td>
</tr>
<tr>
<td>Interruptions</td>
<td>-0.046+</td>
<td>-0.069**</td>
</tr>
</tbody>
</table>

n/s = not significant; * p<.05, ** p<.01, *** p<.001, **** p<.0001
Mixed changes in nonstandardized behaviors as interviewers gain within-study experience across the studies.
Summary: Interviewer behaviors

• Interviewers do change their behaviors as they gain experience

• Interviewers become more efficient in administering questions.
  • Have fewer questions with stutters, disfluencies, and laughter

• Interviewers experience changes in both standardized and non-standardized behaviors, although these replicate less well across studies.
  • In WLT1, few changes in standardized behaviors. In WLT2, fewer standardized behaviors.
  • Across both studies, lose inadequate verification. Other changes in nonstandardized behaviors less consistent.
This paper

• What interviewer behaviors change over the course of the data collection period in two telephone surveys?

• Do these behaviors account for changes in survey length over the course of the data collection period?
What behaviors are associated with overall interview length?

\[
\text{Length}_{ij} = \gamma_{00} + \beta_1 \ln(\text{IwOrder})_{ij} + \beta_2 \text{IwBehaviors}_{ij} + \beta_p \text{Controls}_{pij} + u_{0j} + \varepsilon_{ij}
\]

• Look at the interview order coefficient as groups of behaviors are included in the model.
The interviewer behaviors partially explain interview length. Especially efficiency behaviors in WLT1.

<table>
<thead>
<tr>
<th>behaviors</th>
<th>No behaviors</th>
<th>Including standardized behaviors</th>
<th>Including efficiency behaviors</th>
<th>Including nonstandardized behaviors</th>
<th>Including all behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order Coefficient Predicting Interview Length</td>
<td>-0.55</td>
<td>-0.65 -0.69</td>
<td>-0.31</td>
<td>-0.45</td>
<td>-0.46</td>
</tr>
<tr>
<td></td>
<td>-0.85</td>
<td></td>
<td>-0.71</td>
<td>-0.73</td>
<td>-0.60</td>
</tr>
</tbody>
</table>

WLT1 | WLT2
Takeaways

• Interviewers generally don’t lose their standardized behaviors over the field period.
  • This is good news. Where there are notable losses in standardized behaviors, it appears to be in feedback behaviors (ok; thank you).
  • Standardized behaviors explain between none and 20% of the change in interview length.

• Interviewers do become more efficient in administering surveys over the field period.
  • Efficiency behaviors explain between 17% and all of the change in interview length.

• Interviewers do change in their use of nonstandardized behaviors.
  • Some nonstandardized behaviors (inadequate verification) decrease. May be tradeoffs between major changes in question wording and directive probes.
  • Nonstandardized behaviors explain between 14 and 18% of the change in interview length.
Limitations

• Looked only at interviewer behaviors, but many interviewer behaviors occur in reaction to respondent behaviors.
  • Future research will examine changes in respondent behaviors as well.

• Two surveys conducted two years apart, but one organization conducting the survey.
  • Future research will add in a survey conducted by a different organization.

• Results largely replicate using turns rather than questions.
  • But some model sensitivity to the collection of behaviors included.
Summary

• Interviewer behaviors do change over the course of the data collection period.

• Interviewer behaviors are related to interview length.

• But how interviewer behaviors are related to interview length is more complicated than simply the number of questions on which the behaviors occur over the interview.
  • Are behaviors getting shortened as well as eliminated?
  • How do question characteristics themselves affect the occurrence of these behaviors?
  • More work to be done!
Thanks!

Kristen Olson
@olson_km
kolson5@unl.edu