On the Utility of Paradata in Major National Surveys: Challenges and Benefits

Brady T. West
Survey Research Center, Institute for Social Research
University of Michigan-Ann Arbor
bwest@umich.edu

Frauke Kreuter
Joint Program in Survey Methodology
University of Maryland-College Park
fkreuter@survey.umd.edu

2016 International Methodology Symposium (Statistics Canada)
Presentation Overview

• A summary of research conducted examining the quality and utility of paradata in the U.S. National Survey of Family Growth (NSFG)

• Research on the utility of paradata in major panel surveys in the U.S. and Germany:
  – The Medical Expenditure Panel Survey (MEPS)
  – The Labor Market and Social Security (PASS) study
The NSFG

• The major national fertility survey in the United States
• An important source of data on sexual activity, sexual behavior, and reproductive health for policy makers
• Target population (Until Sept. 2015): Ages 15-44
• Target population (Sept. 2015-Present): Ages 15-49
• Continuous sample design: Four national quarter samples are released and worked each year
• Face-to-face interviews (60-80 minutes) with one person from each household; ACASI for sensitive items
Paradata in the NSFG

• Interviewer Observations
  – Segment (Area) Level (e.g., safety concerns?)
  – Housing Unit Level (e.g., young children present?)
  – Respondent Level (e.g., is selected R sexually active?)
  – Post-survey observations (e.g., ACASI behaviors?)

• Call Record Data
  – Number of call attempts
  – Evidence of refusals, concerns, etc.
Paradata in the NSFG

• Case Disposition Outcomes
  – Respondent, Final Refusal, Non-Sample, etc.

• Keystroke Information
  – Interviewer requests for help, backing up, etc.
  – Respondents changing answers
  – Response timing for individual survey items
The MEPS

• The U.S. Medical Expenditure Panel Survey
• Face-to-face subsample from National Health Interview Survey with five rounds over 2 years
• Paradata:
  – Call Record Data
    • Number of call attempts
    • Evidence of refusals, concerns, etc.
  – Case Disposition Outcomes
    • Respondent, Final Refusal, Non-Sample, etc.
The PASS

• The PASS ‘Labour Market and Social Security’ Study, in Germany

• An annual mixed-mode household survey based on two random samples (welfare benefit recipients and households from a residential building survey); now in 10th panel wave

• Paradata:
  – Call Record Data
  – Case Disposition Outcomes
  – Interviewer Observations (experimental study)
Utility #1: Nonresponse Adjustment

• Interviewer observations collected on all sampled units are included in models of response propensity, which are used to adjust weights
• Observations related to both key outcomes and response propensity have the ability to reduce nonresponse bias

NSFG (West, 2013a)
Current sexual activity of selected R
Presence of young children
Physical impediments to housing units

PASS (West et al., 2014)
Income Bracket (low, med., high)
Anyone in HH on welfare benefits
Challenge #1: Observation Quality

• What if the observations are prone to error?
• They are (West, 2013a; West et al., 2014):
  – Sexual activity: 78% “accuracy”
  – Young children: 72% “accuracy”
  – Benefit receipt: 78% accuracy
  – Income Bracket: 55% accuracy
  – Accuracy also varies substantially among interviewers
    (West and Kreuter, 2013; Sinibaldi et al., 2013; West et al., 2014); Why?
• Error-prone observations will hinder the effectiveness of nonresponse adjustments (West, 2013a; West, 2013b)
Challenge #1: Observation Quality

Figure 1: Variance in Observation Accuracy Among 94 NSFG Interviewers.
2016 International Methodology Symposium (Statistics Canada)
Challenge #1: Observation Quality

<table>
<thead>
<tr>
<th>Interviewer Observed</th>
<th>Unemployment benefit:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-reported</td>
<td></td>
</tr>
<tr>
<td></td>
<td>On UB</td>
<td>Not on UB</td>
</tr>
<tr>
<td>On UB (n=1906)</td>
<td>72.8%</td>
<td>27.2%</td>
</tr>
<tr>
<td>Not on UB (n=1234)</td>
<td>21.9%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Missing (n=73)</td>
<td>43.8%</td>
<td>56.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income:</th>
<th>Self-reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewer Observed</td>
<td>Low</td>
</tr>
<tr>
<td>n=1961</td>
<td>n=684</td>
</tr>
<tr>
<td>Low (n=1511)</td>
<td>82.3%</td>
</tr>
<tr>
<td>Medium (n=1362)</td>
<td>45.2%</td>
</tr>
<tr>
<td>High (n=267)</td>
<td>19.1%</td>
</tr>
<tr>
<td>Missing (n=73)</td>
<td>69.9%</td>
</tr>
</tbody>
</table>
Challenge #1: Observation Quality

• So what can we do about this?
• **One idea:** Provide the interviewers with important predictors of the features they are trying to observe (West and Kreuter, 2015)
• Understand the cues and strategies that different interviewers are using to make their observations, and **standardize training** based on the most effective strategies (West and Kreuter, 2011; West et al., submitted)
Utility #2: Interviewer Evaluation

• The different types of paradata collected inform eligibility, contact (daily), and cooperation (daily) propensity models (Krueger and West, 2014)

• These models are used to compute expectations of ________ propensity at a given point in time

• Interviewer performance can then be evaluated by comparing actual daily outcomes to expectations, and averaging the deviations for a given interviewer (West and Groves, 2013)
Utility #2: Interviewer Evaluation

• Use keystroke information to identify interviewers with unusual tendencies to correct responses or move too quickly
• Intervene with the interviewers in question to improve performance during actual interviews
Utility #2: Interviewer Evaluation

- Factor 1: Too Fast (Z-score based on all items)
- Factor 2: Many Error Checks
- Factor 3: Many ‘Don’t Know’ and ‘Refused’

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Average of Zscore</th>
<th>W08</th>
<th>W10</th>
<th>W12</th>
<th>W06</th>
<th>W08</th>
<th>W10</th>
<th>W12</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor1</td>
<td></td>
<td>-0.36</td>
<td>-0.52</td>
<td>-0.54</td>
<td>-0.79</td>
<td>-0.83</td>
<td>-0.74</td>
<td>-0.82</td>
</tr>
<tr>
<td>factor2</td>
<td></td>
<td>3.48</td>
<td>3.41</td>
<td>3.25</td>
<td>-0.83</td>
<td>-0.36</td>
<td>0.26</td>
<td>0.16</td>
</tr>
<tr>
<td>factor3</td>
<td></td>
<td>0.53</td>
<td>0.74</td>
<td>1.47</td>
<td>0.08</td>
<td>0.50</td>
<td>1.20</td>
<td>1.44</td>
</tr>
</tbody>
</table>

2016 International Methodology Symposium (Statistics Canada)
Challenge #2: Model Specification

• How do we know if a given type of propensity model has been correctly specified?
• Error-prone interviewer observations can once again play a role...
• Should random interviewer effects be included in the models (so that they are evaluated against themselves)?
• What if paradata are missing?
Utility #3: Response Quality

• Use post-survey observations to identify respondents who may be providing data of poor quality

• Assess interviewer behaviors that may affect responses on sensitive items during ACASI (West and Peytcheva, 2014)
Utility #3: Response Quality

• Interviewers vary substantially in terms of how often they sit close enough to see the screen

• ACASI reports on sensitive behaviors vary as a function of whether the interviewer says that they can see the screen
Utility #3: Response Quality
Utility #3: Response Quality

2+ Occasions Using Marijuana

![Chart showing response quality forAbility to see screen and assistance to respondent.](image-url)
Challenge #3: Post-Survey Observations

• Are post-survey interviewer observations reliable indicators of data quality (Wang et al., 2013)?

• Past literature has shown that these are a function of respondent characteristics (rather than data quality), and there is consistent evidence of interviewer variance in them.

• Can these observations be combined to reliably indicate data quality? Open question!
Utility #4: Responsive Survey Design

• The paradata collected in the NSFG are examined daily in an RSD framework to monitor field production and efficiency

• Interventions are implemented when the paradata suggest that certain processes may be introducing bias or inefficiency

• **Example:** Monitoring response rates across different socio-demographic subgroups, and increasing interviewer focus on subgroups found to be lagging (Wagner et al., 2012)
Utility #4: Responsive Survey Design

Response Rates by Subgroup

Day

RR

2016 International Methodology Symposium (Statistics Canada)
Utility #4: Responsive Survey Design

% Respondents and Nonrespondents Judged to Have Kids by Day

- Green: Year3 Quarter2 % Judged with Kids among All Rs
- Red: Year3 Quarter2 % Judged with Kids among All NRs
- Blue: Year3 Quarter2 % Judged with Kids among All Rs+NRs
Challenge #4: Bias Indicators

- Are response rates in different socio-demographic subgroups the best indicators of nonresponse bias?
- Should we really be monitoring response rate variability among other subgroups more closely related to key outcomes (e.g., presence of children)?
  - What if the subgroup variables (e.g., observations) are error-prone?
- Paradata could inform a variety of possible nonresponse bias indicators (Nishimura et al., forthcoming; see also Krueger and West, 2014)
Utility #5: Call Efficiency

• Paradata at MEPS and PASS were used to tailor contact attempts to “best” times, based on historical data.

• In the MEPS, postcards were sent out indicating a personal visit at the same time as last year (Kreuter et al., 2014).

• In PASS, successful call windows were used to reduce time to first contact and interview in subsequent waves (Kreuter and Mueller, 2015).
Utility #5: Call Efficiency

![Chart showing total actions, in-person actions, and telephone actions across different numbers of actions in prior round, with comparison to treatment and control groups.]
Challenge #5: Call Efficiency

- Using previous wave call information has the potential to bias towards stable respondent units → more research is needed to see if prediction of change can be integrated.
- Prescribing contact strategies can be disruptive for established field procedures; buy-in of field representatives is needed.
- Optimal allocation of contact times is key.
Summary

• These three surveys are committed to using a variety of paradata to improve their operations and their ultimate data products

• An active program of research on paradata is necessary to fully understand (and improve) the measurement error properties of these data

• All three surveys have a long history of collaborating with other researchers interested in these areas of research!
References


References


References


Thank You!

• Please do not hesitate to send any questions to bwest@umich.edu or fkreuter@umd.edu.
• We would welcome questions at this point.