A Comparison Of Ex-ante, Laboratory, and Field Methods for Evaluating Survey Questions

Aaron Maitland, Westat
Background and Motivation

Concepts

Item characteristics

Response process

Researcher

Questionnaire

Respondent

Data Quality

2
Question Evaluation Methods

• **Online computer tools**
  – Question Understanding Aid (QUAID): Online tool that analyzes questions for various linguistic characteristics that may be problematic.
  – Survey Quality Prediction Tool (SQP): Computer tool used to code characteristics of a survey questions and predict data quality.

• **Expert methods**
  – Expert Review: Questionnaire reviewed independently by individual experts.
  – Questionnaire Appraisal System (QAS): Standardized form to check questions for specific problems.

• **Laboratory methods**
  – Cognitive interviewing: Subjects think aloud or cognitive probing techniques are used to understand responses to the questions.

• **Field methods**
  – Behavior coding: Code the interviews for problematic interviewer or respondent behavior.
  – Response latency: Measure time between the end of the reading of a question and beginning of the answer.

• **Reliability and validity**
  – Test-retest reliability
  – Record check
Previous Research

• Study designs
  – Descriptive
    • Describe problems found by different methods
    • No inferences made
  – Exploratory
    • Compare methods on number and type of problems
    • Agreement or correlational statistics are also common
  – Confirmatory
    • Predicts quality of the data in the field
    • Dependent variables are either indirect assessments of quality in the field such as behavior coding results or direct assessments such as reliability or validity
Previous Research

• Studies have produced inconsistent findings (Presser and Blair, 1994; Rothgeb, Willis, and Forsyth, 2001; Willis, Schechter, and Whitaker, 1999; Yan, Kreuter, and Tourangeau, 2012).

• Very little research on how well pretest methods predict data quality (Dykema, Lepkowski, and Blixt, 1997; Hess, Singer, and Bushery, 1999; Yan, Kreuter, and Tourangeau, 2012).

• When data quality is studied, usually only one method is studied (Dykema, Lepkowski, and Blixt, 1997; Hess, Singer, and Bushery, 1999).
Outstanding Issues

• What are the circumstances in which the methods agree?

• How can we package the methods together to identify problems that affect data quality?

• What is the relative effectiveness of the methods?
Hypotheses

• Complementary methods hypothesis
  – Using multiple methods to predict problems or reliability is better than using a single method.

• Test environment hypothesis
  – Methods implemented in a more realistic setting should be most closely related to data quality.
Data Sources

• Primary data come from the 2006 JPSM survey practicum.
  – Questionnaires included 88 factual and attitudinal questions.
  – Interviews for the final study were recorded.
  – Study included a reinterview of 53 questions two weeks after the initial interview.

• Supplementary questions from record check studies from the literature.
Data Preparation: Qualitative Coding for Ex-ante and Laboratory Methods

- Semantic I: Problems with question structure
  - Information overload, sentence structure, transition problem
- Semantic II: Problems with meaning of terms
  - Vagueness or ambiguity, technical or common terms not understood, double-barreled
- Respondent task I: Problems with recall
  - Recall is difficult, response is resisted
- Respondent task II: Problems with response categories
  - Overlapping or insufficient response categories, response categories not appropriate
- Respondent task III: Problems with sensitivity
  - Social desirability
- Other problems
  - Procedural issues, analysis issues
Data Preparation: Field Methods

• Behavior coding
  – Interviewer codes
    • Exact reading, slight change, major change, probing
  – Respondent codes
    • Adequate answers, qualified answers, inadequate answers, don’t know, refusal, respondent interruption, pauses or fillers

• Response latency:
  – Used recordings to measure time from the end of the reading of the question by the interviewer to the onset of the answer by the respondent
## Prediction of the Accuracy of Questions

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.87**(.02)</td>
<td>.88**(.02)</td>
<td>.88**(.02)</td>
<td>.85**(.03)</td>
<td>.81**(.02)</td>
</tr>
<tr>
<td>Expert review</td>
<td>-.10**(.02)</td>
<td>-.09**(.02)</td>
<td>-.09**(.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAS</td>
<td>.02(.02)</td>
<td></td>
<td></td>
<td>-.05**(.02)</td>
<td></td>
</tr>
<tr>
<td>Cognitive interviewing</td>
<td>-.28**(.12)</td>
<td>-.24**(.11)</td>
<td></td>
<td></td>
<td>-.26*(.14)</td>
</tr>
<tr>
<td>R-squared</td>
<td>.41</td>
<td>.41</td>
<td>.36</td>
<td>.10</td>
<td>.07</td>
</tr>
</tbody>
</table>

**p<.05, *p<.10**
Prediction of Field Results

• Dependent variables: behavior coding, response latency, item nonresponse

• Independent variables: results from QUAID, SQP, expert review, QAS, and cognitive interviewing

• Complementary methods hypothesis:
  – Compare full and reduced models. Can the model be reduced to a single method?

• Test environment hypothesis – Methods implemented in a more realistic setting should be most closely related to data quality.
  – Compare model fit across models (AIC)
Prediction of the Level of Problems (Multilevel Models)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Adequate Answers</th>
<th>Requests for Clarification</th>
<th>Item Nonresponse</th>
<th>Response Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complementary methods</td>
<td>Supported</td>
<td>Supported</td>
<td>Supported</td>
<td>Not Supported</td>
</tr>
<tr>
<td><strong>Predictors in reduced model</strong></td>
<td>QAS sensitivity problems and Cognitive interviewing recall problems</td>
<td>Expert Review, QAS, Cognitive Interviewing problems with question meaning</td>
<td>QAS response category problems and Cognitive interviewing recall problems</td>
<td>QAS recall problems</td>
</tr>
<tr>
<td>Test environment</td>
<td>Partially supported</td>
<td>Not supported</td>
<td>Supported</td>
<td>Not supported</td>
</tr>
<tr>
<td><strong>ordering of methods</strong></td>
<td>CI and QAS &gt; ER and SQP</td>
<td>CI and ER and QAS &gt; QUAID</td>
<td>CI &gt; QAS &gt; ER &gt; SQP and QUAID</td>
<td>QAS &gt; CI and ER and QUAID &gt; SQP</td>
</tr>
</tbody>
</table>
Prediction of Reliability

• Reinterview occurred two weeks after first interview
• Question level models predicted the index of inconsistency using all available methods.
• SQP total quality (for subjective questions only), QAS recall problems, cognitive interviewing recall problems, QUAID response category problems, and expert review sensitivity issues gave best predictions of reliability.
Conclusions

• Usually necessary to use multiple methods to get the best predictions.

• Direct observation of response process is not always necessary to detect problems with questions, however, cognitive interviewing is often among the best predictors of poorly performing questions.

• Further research needed to understand how to fit the results from online computer tools with traditional method evaluations.

• Future research should be focused on evaluating multiple methods using questions with known psychometric properties.