Individual Ability Estimation & Classification Accuracy Under Rapid Guessing Misidentifications

Joseph A. Rios

University of Minnesota

Presentation at the annual conference of the National Council on Measurement in Education 4/23/22
Handling Rapid Guessing

• Two general scoring approaches when leveraging response times as a proxy of RG: (a) mixture models; (b) response time threshold scoring

• Response time threshold scoring consists of two steps:
  – Establish a threshold that differentiates between solution behavior and RG
  – Rescore all RG responses as incorrect (Deribo et al., 2021; Wright, 2016) or missing (Liu et al., 2019; Wise & DeMars, 2006)

• Assumptions of these rescoring approaches:
  – Incorrect scoring: RG occurs due to item difficulty
  – Missing scoring: RG occurs completely at random
Effort-Moderated (EM) Scoring

• EM scoring assumes that any response classified as RG is an invalid indicator of examinee ability
  – RG responses are classified using response time thresholds
    • These responses are treated as missing data
  – Probabilities for all other responses are estimated based on a traditional IRT model

• EM scoring is one of the most popular approaches used in research and practice, as it is:
  – Flexible (i.e., does not assume a strict form of rapid guessing pattern)
  – Computationally simple
Rationale for Current Study

• Nearly all studies to date have assumed that RG is perfectly identified, an underlying assumption of all response-time threshold scoring approaches
  – The tenability of this assumption is unknowable in operational settings due to the use of log file information to proxy examinee behavior

• Little is known about the performance of the EM scoring when RG is imperfectly classified and individual scores are reported
  – Potential for RG in formative assessments (e.g., Wise & Kingsbury, 2016)
Research Questions

1. What is the ability estimate bias difference between EM scoring when RG is perfectly and imperfectly identified at different ability levels?

2. To what extent are incorrect RG identifications associated with classification errors of ability?
METHOD
Data Generation of Effortful Response Probabilities

- Data were generated for a 50-item test using the unidimensional 2PL model

- Five discrete simulee abilities were sampled: -2, -1, 0, 1, 2

- Generating item parameters were taken from a NAEP math test
  - Mean discrimination: 1.05 ($SD = 0.35$)
  - Mean difficulty: 0.07 ($SD = 0.67$)
## Manipulation of RG Misclassifications

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>% RG Responses</td>
<td>10%, 20%, 30%, 40%</td>
</tr>
<tr>
<td>Misclassification Type</td>
<td>Underclassification or Overclassification</td>
</tr>
<tr>
<td>Rate of Misclassifications</td>
<td>20%, 40%, 60%</td>
</tr>
</tbody>
</table>

All independent variables were fully-crossed (24 conditions) Each condition was replicated 10,000 times at each ability level
Analyses

• Ability parameters were estimated based on two scoring approaches using fixed item parameter estimation:
  – (a) a 2PL model that ignores the presence of RG responses
  – (b) EM scoring which downweights all RG responses to have zero contributions to estimation

• Ability parameters were obtained via ML estimation

• Bias and classification (cut-score of 0.46 logits) accuracy were calculated
RESULTS
Ability Parameter Estimate Accuracy
Classification Accuracy
DISCUSSION
Summary

- Across conditions, bias tended to be on average lower when RG was overclassified as opposed to underidentified.

- Although EM scoring generally reduced bias introduced by RG, it was susceptible to classification inaccuracies:
  - Driven by increased standard errors from reducing the amount of information available in ability estimation.
Limitations & Future Research Directions

- Only compared two scoring approaches that were selected based on their computational simplicity and popularity in the literature
  - Comparison of EM-IRT to mixture-model and robust ML approaches

- Fixed item parameter MLE was employed
  - Assumed accurate item parameter estimates are available, which may be untenable for some testing programs

- Classification errors were examined for a testing context in which test information was not maximized around the cut-score
  - Though, mean difficulty ($\bar{b} = 0.07$) for the simulated test did not drastically differ from the cut-score (0.46 logits)
Implications

• It is better to incorrectly classify RG, then to ignore such responses

• When choosing a response time threshold to identify RG, practitioners may benefit from adopting liberal thresholds to limit false negative classifications of RG
Thank You

For any questions, please feel free to contact me:

jrios@umn.edu

Slides for this presentation can be found on my personal website:
https://josephriosphd.com/