Development of Multistage Tests based on Teacher Ratings

Stéphanie Berger\textsuperscript{12}, Jeannette Oostlander\textsuperscript{1}, Angela Verschoor\textsuperscript{3}, Theo Eggen\textsuperscript{23} & Urs Moser\textsuperscript{1}

1 Institute for Educational Evaluation, Associated Institute of the University of Zurich
2 Research Center for Examinations and Certification, University of Twente
3 CITO

IACAT Cambridge, 14\textsuperscript{th} to 16\textsuperscript{th} of September 2015
Overview

• Introduction
• Test development
  – Test design
  – Test construction based on teacher ratings
  – Routing rules based on heuristics
• Results
  – Routing
  – Correlation between ratings and item difficulty
  – Information per module and path
  – Reliability
• Discussion and conclusion
Introduction

• Development of standardized tests for secondary school in Northwestern Switzerland

• Assessment of student ability in four different school subjects
  – Individual reporting
  – High stakes

• Target population:
  – Secondary school students, grade 8
  – Three different school types

• Content framework: New Swiss curriculum

• Computer-based assessment
Research Question

• Target population covering a broad ability range → Multistage testing
• New item pool, but no resources for pretesting → Teacher ratings as approximation of item difficulty

Questions
• What are the implications of using teacher ratings instead of pretest data for constructing a multistage test?
• Do teacher ratings allow us to construct a reliable multistage test?
Advantages of Multistage Testing
Yan, Lewis & von Davier (2014)

• Adaptive optimization of fit between item difficulty and student ability
• More efficient and precise measurement of student ability compared to linear tests
• Higher control over content balance and test structure compared to fully adaptive tests
• Allows students to navigate and to review items within one module
• Reduced test copying compared to linear tests
Multistage Test Design Mathematics

Practical considerations:

- Testing time: 2 lessons = 90 minutes
- Reduce copying by multiple versions
- Allow for recovery from inadequate routing

→ “Double“ 1-3-3-3 MST including 252 items
Test Construction based on Teacher Ratings

Teacher ratings of item difficulty

- 6 secondary school teachers from Northwestern Switzerland
- Rating of printed items including item key
- Categorization of items into three different categories: easy, medium, difficult
Distribution of Items per Module
Routing Rules based on Heuristics

- Routing based on raw score
- Target difficulty per module: $p = 0.66$
  - Predicted mean score: $2/3$ of maximum score
  - Predicted SD: $1/6$ of maximum score
- Goal to route equal amount of students per path
  - $\frac{1}{3}$ per path for routing module and medium modules
    → routing based on $P_{33}$ and $P_{66}$ of predicted score
  - $\frac{1}{2}$ per path for easy and difficult modules
    → routing based on mean of predicted score
Routing Rules based on Heuristics

Max = 9
→ x = 6.0, SD = 1.5
→ P_{33} = 5.3, P_{66} = 6.6

Max = 14
→ x = 9.3, SD = 2.3

Max = 16
→ X = 10.7, SD = 9.5
→ P_{33} = 9.5, P_{66} = 11.8

…
Calibration

• Sample: $N = 7176$ grade 8 students
• Item response model: One Parameter Logistic Model (OPLM) (Verhelst & Glas, 1995)
  
  $$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp[a_i(\theta_i - \beta_j)]}{1 + \exp[a_i(\theta_i - \beta_j)]}$$

• Item calibration with OPLM program (Verhelst, Glas & Verstralen, 1995)
• Marginal maximum likelihood estimation (MML)
• Exclusion of 15 items due to poor model fit, low discrimination or low p-value
## Results I: Descriptive Values per Module

<table>
<thead>
<tr>
<th>St.</th>
<th>Module</th>
<th>Lev.</th>
<th># Items</th>
<th>Mean β</th>
<th>Mean SE(β)</th>
<th># Observations</th>
<th>% Observations</th>
<th>Mean θ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1A</td>
<td>R</td>
<td>8</td>
<td>-1.041</td>
<td>.033</td>
<td>3659</td>
<td>51%</td>
<td>-0.538</td>
</tr>
<tr>
<td></td>
<td>1B</td>
<td>R</td>
<td>8</td>
<td>-.988</td>
<td>.049</td>
<td>3518</td>
<td>49%</td>
<td>-0.512</td>
</tr>
<tr>
<td>2</td>
<td>2A</td>
<td>E</td>
<td>9</td>
<td>-.913</td>
<td>.051</td>
<td>1810</td>
<td>25%</td>
<td>-1.135</td>
</tr>
<tr>
<td></td>
<td>2B</td>
<td>E</td>
<td>9</td>
<td>-.354</td>
<td>.059</td>
<td>1811</td>
<td>25%</td>
<td>-1.162</td>
</tr>
<tr>
<td>3</td>
<td>3A</td>
<td>M</td>
<td>8</td>
<td>.390</td>
<td>.060</td>
<td>1239</td>
<td>17%</td>
<td>-0.108</td>
</tr>
<tr>
<td>3B</td>
<td>M</td>
<td>9</td>
<td>.132</td>
<td>.067</td>
<td>1099</td>
<td></td>
<td></td>
<td>-0.100</td>
</tr>
<tr>
<td>4</td>
<td>4A</td>
<td>D</td>
<td>8</td>
<td>1.766</td>
<td>.161</td>
<td>588</td>
<td>8%</td>
<td>0.546</td>
</tr>
<tr>
<td>4B</td>
<td>D</td>
<td>7</td>
<td>.557</td>
<td>.215</td>
<td>628</td>
<td></td>
<td></td>
<td>0.501</td>
</tr>
<tr>
<td>5</td>
<td>5A</td>
<td>E</td>
<td>15</td>
<td>-.480</td>
<td>.076</td>
<td>1969</td>
<td>27%</td>
<td>-1.156</td>
</tr>
<tr>
<td>5B</td>
<td>E</td>
<td>14</td>
<td>-.207</td>
<td>.073</td>
<td>1884</td>
<td></td>
<td></td>
<td>-1.144</td>
</tr>
<tr>
<td>6</td>
<td>6A</td>
<td>M</td>
<td>15</td>
<td>-.157</td>
<td>.064</td>
<td>1348</td>
<td>19%</td>
<td>0.042</td>
</tr>
<tr>
<td>6B</td>
<td>M</td>
<td>14</td>
<td>-.520</td>
<td>.076</td>
<td>1302</td>
<td></td>
<td></td>
<td>0.041</td>
</tr>
<tr>
<td>7</td>
<td>7A</td>
<td>D</td>
<td>15</td>
<td>.364</td>
<td>.117</td>
<td>328</td>
<td>5%</td>
<td>0.892</td>
</tr>
<tr>
<td>7B</td>
<td>D</td>
<td>13</td>
<td>1.478</td>
<td>.192</td>
<td>322</td>
<td></td>
<td></td>
<td>0.813</td>
</tr>
<tr>
<td>8</td>
<td>8A</td>
<td>E</td>
<td>13</td>
<td>-1.070</td>
<td>.057</td>
<td>2052</td>
<td>29%</td>
<td>-1.123</td>
</tr>
<tr>
<td>8B</td>
<td>E</td>
<td>13</td>
<td>-.323</td>
<td>.067</td>
<td>2087</td>
<td></td>
<td></td>
<td>-1.103</td>
</tr>
<tr>
<td>9</td>
<td>9A</td>
<td>M</td>
<td>15</td>
<td>-.143</td>
<td>.067</td>
<td>1346</td>
<td>19%</td>
<td>0.128</td>
</tr>
<tr>
<td>9B</td>
<td>M</td>
<td>14</td>
<td>-.273</td>
<td>.074</td>
<td>1239</td>
<td></td>
<td></td>
<td>0.098</td>
</tr>
<tr>
<td>10</td>
<td>10A</td>
<td>D</td>
<td>15</td>
<td>1.021</td>
<td>.172</td>
<td>296</td>
<td>4%</td>
<td>1.007</td>
</tr>
<tr>
<td>10B</td>
<td>D</td>
<td>15</td>
<td>.607</td>
<td>.170</td>
<td>298</td>
<td></td>
<td></td>
<td>0.914</td>
</tr>
</tbody>
</table>
Results II: Routing

from 1A/B

from 2A/B

from 3A/B

from 4A/B

from 5A/B

from 6A/B

from 7A/B
Result III: Correlation between Ratings and Item Difficulty

$r = 0.44$
$n = 220$
$p < 0.01$
Results III: Information per Module

Stage 1 - Routing

Stage 3

Stage 2

Stage 4
Results IV: Information per Path

The graph illustrates the information per path for different ability levels, categorized by difficulty. The x-axis represents ability (θ) ranging from -7 to 7, and the y-axis shows information per path (I(θ)) ranging from 0 to 30. The curves are color-coded to indicate different difficulty levels:

- Easy (n = 2573)
- Medium (n = 551)
- Difficult (n = 119)
- Easy-Diff (n = 62)
- Diff-Easy (n = 218)

Each curve shows how information per path varies with ability for each difficulty category.
Results V: Test Reliability

Simulation

- Item parameters from calibration
- 50,000 simulees from $N(\text{mean} = -0.546, \text{SD} = 0.890)$

Estimated reliability: $\rho = \frac{\text{Var}(T)}{\text{Var}(X)} = \frac{\text{Var}(\theta)}{\text{Var}(\bar{\theta})}$

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Mean test length</th>
<th>Mean test score</th>
<th>Estimated reliability</th>
<th>Test length comp. rel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multistage test</td>
<td>44.9</td>
<td>22.0</td>
<td>0.90</td>
<td>35.8</td>
</tr>
<tr>
<td>Random linear test</td>
<td>45.0</td>
<td>18.5</td>
<td>0.87</td>
<td>56.5</td>
</tr>
</tbody>
</table>
Discussion & Conclusion

• Moderate correlation between teacher ratings and estimated item difficulty
• General underestimation of item difficulty
• Multistage item collection designs involve risk of unbalanced number of observations per module
• Higher reliability of multistage test compared to a random linear test
Questions and Discussion

Contact: Stephanie.Berger@ibe.uzh.ch
References I

