Development of an Online Item Bank for Adaptive Formative Assessment

Stéphanie Berger¹², Urs Moser¹, Angela Verschoor³ & Theo Eggen²³

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

AEA-Europe Conference, Glasgow, 5th to 7th November 2015
Outline

• Introduction:
  – Educational assessment in Northwestern Switzerland
  – Formative assessment
  – Research Questions

• Method:
  – Online item bank for adaptive formative assessment
  – Development of a vertical IRT scale
  – Efficient data collection designs

• Selected Results

• Conclusion
INTRODUCTION
Educational assessment in Northwestern Switzerland

- Initiative of four cantons in Northwestern Switzerland
- N = 13’000 students per school year
- Instruments:
  - Four compulsory standardized tests
  - Online item bank for formative assessment
- Assessment of student ability in five different school subjects
  - German (first language)
  - English, French
  - Mathematics
  - Science
Standardized Tests and Formative Assessments

- Standardized Tests
  - Stand. Test Grade 3
  - Stand. Test Grade 6
  - Stand. Test Grade 8
  - Stand. Test Grade 9

- Formative Assessments

School Grades:
3 4 5 6 7 8 9
Formative Assessment
Van der Kleij, Vermeulen, Schildkamp & Eggen, 2015

• Assessment to support learning and to accommodate students’ individual educational needs
• Focus on student and class level
• Three different approaches
  – Data-based decision making
  – Assessment for Learning
  – Diagnostic testing
Data-based Decision Making (DBDM)
Schildkamp & Kuiper, 2010; Van der Kleij, Vermeulen, Schildkamp & Eggen, 2015

• Focus on learning outcome
• Systematic collection and analysis of objective data
• Feedback loop
  – Assessment of students’ current ability level
  – Definition of appropriate learning goals based on assessment outcomes
  – Monitoring/evaluation of students’ progress through further assessments
  – Adjustment of goals and/or learning environment based on the assessment outcomes
Research Questions

• Students differ in their current ability level
• Students follow their individual learning goals
• Students differ in their learning progress

→ How can we provide students and teachers tailored assessments that consider the individual differences?
→ Which psychometrical challenges are related to the development of such an assessment system?
METHOD
Online Item Bank for Adaptive Formative Assessment

• Online item bank
  – Huge amount of assessment items that cover several school grades
  – Access through web browsers → on demand administration

• Adaptive
  – Assessment administration based on algorithm for computer adaptive testing (CAT) (e.g., Wainer. 2000)
  – Item response theory (IRT) (e.g., Wainer & Mislevy, 2000)
    → Student ability and item difficulty on the same scale
    → Student ability can be estimated on the same scale with any subset of items of the calibrated item bank
CAT: Optimization of Fit of Item Difficulty and Student Ability

- Ability $\theta$ / Difficulty $\beta$
- Confidence interval of $\theta$
- Wrongly answered item
- Correctly answered item

Number of administered items
Psychometrical Challenges

• Development of a vertical IRT scale over several grades
  → Compare abilities over different school grades and measure progress
Development of a Vertical IRT Scale: Concept

- Compare results over time → progress reporting
- Criterion referenced interpretation → meaningful scores
Development of a Vertical IRT Scale: Linking

Common-item non equivalent group design (Kolen & Brennan, 2014)

- Administer identical items to students from neighboring grades
- Estimate item difficulty on one scale using IRT methods

<table>
<thead>
<tr>
<th>Item blocks</th>
<th>Populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td>Grade 3</td>
</tr>
<tr>
<td>Grade 3-4</td>
<td>Grade 4</td>
</tr>
<tr>
<td>Grade 4</td>
<td>Grade 5</td>
</tr>
<tr>
<td>Grade 4-5</td>
<td>Grade 6</td>
</tr>
<tr>
<td>Grade 5</td>
<td>…</td>
</tr>
<tr>
<td>Grade 5-6</td>
<td>…</td>
</tr>
<tr>
<td>Grade 6</td>
<td>…</td>
</tr>
</tbody>
</table>

Legend:
- Blue: Grade specific items
- Yellow: Link items over neighboring school grades
Psychometrical Challenges

• Development of a vertical IRT scale over several grades
  → Compare abilities over different school grades and measure progress

• Efficient data collection design to calibrate the item pool
  → Information about item difficulty needed for adaptive item selection in CAT
Efficient data Collection Design: Linear vs. Multistage

- Limited resources
- Limited number of participants

Linear design for targeted testing

<table>
<thead>
<tr>
<th>Item blocks</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 3-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 5-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multistage data collection design

<table>
<thead>
<tr>
<th>Populations</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Efficient Data Collection Design: Simulation Study

- Samples: $n_1(0.0, 1.0) = 1'300$, $n_2(0.8, 1.0) = 1'300$
- Item pool: 180 dichotomous items, $-2.0 \leq b \leq 2.8$
- Test length: 30 items per student
- Two designs:
  - Linear design
  - Multistage design
- Two conditions:
  - Optimal distribution of items over booklets/modules
  - Distribution based on limited information $\rightarrow$ dislocated items
Efficient Data Collection Design: Designs in Detail

Linear Design
→ 6 equally difficult booklets per sample

Multistage Design
→ 12 paths per sample
(6 easy & 6 difficult paths)
SELECTED RESULTS
Efficient Data Collection Design: Results I

![Graph showing optimal designs with RMSE(b) on y-axis and b on x-axis, with two lines labeled Lin Easy / Diff and Lin Medium.](image)
Efficient Data Collection Design: Results I

Optimal Designs

- Lin Easy / Diff
- Lin Medium
- MS VEasy / VDiff
- MS Easy / Diff
- MS Medium
Efficient Data Collection Design: Results II

Designs with dislocated items ($r = 0.60$)

- Lin Easy / Diff
- Lin Medium
Efficient Data Collection Design: Results II

Designs with dislocated items ($r = 0.60$)
CONCLUSION
Conclusion

- Computer adaptive formative assessment
  → Tailored assessments, high measurement efficiency
  → Automated test creation, administration and scoring
- Vertical IRT scale
  → Compare assessment results over time (cross sectional and longitudinal comparison)
- Calibration of the item bank
  → Linear design for targeted testing
Questions and Discussion

Contact: Stephanie.Berger@ibe.uzh.ch
         Angela.Verschoor@cito.nl
References I


References II

