Examining Item and Testlet Position Effects in Computer-Based Alternate Assessments

Okan Bulut
University of Alberta

Xiaodong Hou
American Institutes for Research

Ming Lei
American Institutes for Research

NCME – April 11, 2016
Why Different Item Positions?

- In computer-based assessments, positions of items are typically scrambled across students to reduce the possibility of cheating among students.

- This practice is highly effective for ensuring the integrity of assessments; however, varying item positions significantly may result in position effects.

- Position effects may appear due to
  - Fatigue at the end of the test
  - Test wiseness because of familiarity with test content
  - Lack of motivation towards the end of the test
Impact of Item Positions

- Position effects are usually ignored because of
  - item invariance assumption in IRT, and
  - the assumption that every student is affected equally by position change.

- Especially with field-test items, these assumptions may not hold.
  - Students can respond to field-test items in different positions.

- When field-test items become operational, they may need to be used in positions different from the positions that they were originally calibrated.

- If computerized-adaptive testing is used, there is no way of controlling positions of field-test items.
Studies on Position Effects

- TIMSS (Martin, Mullis, Gonzalez & Chrostowski, 2004)

- PISA (Debeer & Janssen, 2013)

- GRE (Albano, 2013; Davey & Lee, 2011)

- Achievement tests from Germany and Austria (Hohensinn et al., 2011; Robitzsch, 2009)

- A high-stakes reading assessment was used.

- DIF between students with disabilities and student without disabilities.

- Students with disabilities consistently underperformed on the second half of the items relative to the first half of the items.

- The number of DIF items increases in the second half of the test.

- Students with disabilities might not have sufficient time or energy to complete the test.
Purpose of Our Study

This study aims to examine position effects in mathematics items in a computer-based alternate assessment.

1. Do mathematics items in the alternate assessment exhibit any significant position effects?

2. Do groups of mathematics items (i.e., testlets) exhibit any significant position effects?

3. What is the relationship between ability levels of students with disabilities and item/testlet position effects?
Data Source

- The data for this study come from the 2014 Spring administration of a statewide alternate assessment program.

- The test has been designed for students with disabilities who have a single disability or multiple disabilities.

- The test administered across three grade bands:
  1. Grades 3 through 5 (N=2465)
  2. Grades 6 through 8 (N=2561)
  3. High School (HS) (N=858)

- The sample of this study included the students who took the mathematics test.
Test Design

- The test consists of twelve operational tasks (i.e., testlets) and one field-test task.

- Each test includes four to eight polytomously-scored items related to a common stimulus.

- The items become increasingly more complex and difficult within a task.

- The tasks become increasingly more complex as the student moves from the initial tasks toward the final tasks on the test.
A semi-adaptive approach was used.
Initial ability levels were determined by SPQ.

- **Group 1 (low ability):** Tasks 1 through 5 + **field-test task**
- **Group 2 (medium ability):** Tasks 3 through 9 + **field-test task**
- **Group 3 (high ability):** Tasks 6 through 12 + **field-test task**
Stopping Rule

- Each student has to complete the minimum number of tasks determined based upon their initial ability level.

- Once a student finishes the required tasks, he/she can continue the test as long as he/she can obtain a minimum of 6 points from the task.

- The student exits the test (i.e., the operational part)
  - If he/she obtains less than 6 points at a task
  - If he/she reaches the last task (i.e., Task 12)

- Then, each student has to complete the field-test task.
Data Analysis

- Position effects were only examined in the field-test task for each grade band because:
  - the field-test task was taken by all students regardless of on which task they started the test and where they exited the test,
  - the field-test task was administered at the end of the test, thereby having higher likelihood of exhibiting position effects due to fatigue; and
  - none of the operational tasks were taken by all students due to different starting points.
Data Analysis

- A structural equation modeling (SEM) approach described by Bulut, Guo, and Gierl (2016) was used to implement three models:
  - Model 0: Partial Credit Model (Masters, 1982) assuming no position effect
  - Model 1: Task position effect
  - Model 2: Item position effect

- The positions of the field-test task and field-test items were determined based on how many operational tasks and items the students were able to complete before moving to the field-test task.
Model 0: Partial Credit Model
Data Analysis

Model 1: Task Position Effect
Data Analysis

Model 2: Item Position Effect
Evaluation Criteria

- Significant parameters for task position and item position effects

- Overall model comparison using
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Sample size adjusted BIC
Results

Grade Band 3-5

% of Students

Field-Task Position

Group 1
Group 2
Group 3
Results

Grade Band 6-8

% of Students

Field-Test Position

Group 1  Group 2  Group 3
Results

Grade Band HS

% of Students

Field-Test Position

- Group 1
- Group 2
- Group 3

Psychometric Issues in Alternate Assessments
**Results**

<table>
<thead>
<tr>
<th>Grade Band</th>
<th>Ability Group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5</td>
<td>G1: Low ability</td>
<td>1356</td>
<td>0.014</td>
<td>1.186</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>G2: Medium ability</td>
<td>735</td>
<td>0.771</td>
<td>0.506</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>G3: High ability</td>
<td>399</td>
<td>1.114</td>
<td>0.546</td>
<td>0.027</td>
</tr>
<tr>
<td>HS</td>
<td>G1: Low ability</td>
<td>544</td>
<td>0.313</td>
<td>1.122</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>G2: Medium ability</td>
<td>199</td>
<td>1.139</td>
<td>0.635</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>G3: High ability</td>
<td>90</td>
<td>1.520</td>
<td>0.582</td>
<td>0.061</td>
</tr>
</tbody>
</table>

**Note:** $SE = SD / \sqrt{N}$. 

*Summary of Estimated Abilities Across Ability Groups and Grade Bands*
## Results

### Table 2

*Summary of Testlet Position Effects for the Field-Test Task*

<table>
<thead>
<tr>
<th>Grade Band</th>
<th>Number of Items</th>
<th>Position Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5</td>
<td>6</td>
<td>0.008 (0.010)</td>
</tr>
<tr>
<td>6-8</td>
<td>6</td>
<td>-0.068 (0.010)**</td>
</tr>
<tr>
<td>HS</td>
<td>6</td>
<td>-0.003 (0.016)</td>
</tr>
</tbody>
</table>

**Note:** Standard errors of the estimated effects are shown in the parentheses; HS: High School; *p < .05, **p < .01, ***p < .001.
## Results

### Table 3

**Summary of Estimated Item Position Effects Across Grade Bands**

<table>
<thead>
<tr>
<th>FT Item</th>
<th>Position Effect</th>
<th>FT Item</th>
<th>Position Effect</th>
<th>FT Item</th>
<th>Position Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>.009 (.003)**</td>
<td>72</td>
<td>-.012 (.005)**</td>
<td>71</td>
<td>.001 (.007)</td>
</tr>
<tr>
<td>70</td>
<td>.006 (.003)</td>
<td>73</td>
<td>-.012 (.004)**</td>
<td>72</td>
<td>.002 (.006)</td>
</tr>
<tr>
<td>71</td>
<td>.004 (.004)</td>
<td>74</td>
<td>-.010 (.003)**</td>
<td>73</td>
<td>-.003 (.005)</td>
</tr>
<tr>
<td>72</td>
<td>-.007 (.004)</td>
<td>75</td>
<td>-.013 (.003)***</td>
<td>74</td>
<td>-.002 (.005)</td>
</tr>
<tr>
<td>73</td>
<td>.006 (.004)</td>
<td>76</td>
<td>-.012 (.003)***</td>
<td>75</td>
<td>-.001 (.005)</td>
</tr>
<tr>
<td>74</td>
<td>-.011 (.003)**</td>
<td>77</td>
<td>-.010 (.003)**</td>
<td>76</td>
<td>-.001 (.005)</td>
</tr>
</tbody>
</table>

**Note:** Standard errors of the estimated effects are shown in the parentheses; HS: High School; FT: Field-Test; *p < .05; **p < .01; ***p < .001;
## Results

**Table 4**

*Summary of Model Fit Statistics from the Three SEM Models Across Grade Bands*

<table>
<thead>
<tr>
<th>Grade Band</th>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>Adjusted BIC</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5</td>
<td>$M_0$: Partial Credit model</td>
<td><strong>234138.734</strong></td>
<td><strong>234952.580</strong></td>
<td><strong>234507.766</strong></td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>$M_1$: Task position effect</td>
<td>234139.798</td>
<td>234959.458</td>
<td>234511.466</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>$M_2$: Item position effect</td>
<td>234122.312</td>
<td>234971.038</td>
<td>234507.160</td>
<td>146</td>
</tr>
<tr>
<td>6-8</td>
<td>$M_0$: Partial Credit model</td>
<td>259143.369</td>
<td>260013.055</td>
<td>259542.816</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td>$M_1$: Task position effect</td>
<td><strong>259080.200</strong></td>
<td><strong>259955.763</strong></td>
<td><strong>259482.346</strong></td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>$M_2$: Item position effect</td>
<td>259089.575</td>
<td>259994.519</td>
<td>259505.215</td>
<td>154</td>
</tr>
</tbody>
</table>

The best fitting model was highlighted in bold in each grade band.

Psychometric Issues in Alternate Assessments
Conclusions & Implications

- Only the field-test task from grade band 6-8 exhibited significant task and item position effects.

- Because the field-test items were not used in scoring, position effects did not have any impact on estimated test scores.

- The next steps should be
  - to monitor these items when they become operational, and
  - to investigate the consequential impact of position effects on students’ proficiency levels.
Thank you!

For further information please contact:
Okan Bulut
bulut@ualberta.ca