Investigating the Efficacy of Link Items in the Construction of an International Mathematics Scale

Rassoul Sadeghi, PhD
University of New South Wales Global
Educational Assessment Australia (EAA)
Nowadays, education authorities place great importance on participating in international assessment programs that are designed to improve student performance.

Such programs provide important information on the achievement of students in a specific country relative to that of students in other countries, and identify factors that have contributed to this achievement.

It is widely recognised that curriculum differences may have a significant impact on equivalence of test items in international assessment programs.

As a result, determining the cross-country equivalence of international test items is a challenge.
Background

- There are different statistical procedures to identify differentially functioning test items and tackle them before performing test equating and scaling to make sure assessments are fair for all examinees.

- Differential Item Functioning (DIF) in psychometric tests has long been recognized as a potential source of bias in person measurement.

- Thus if there is any evidence of systematic differential item functioning by gender/ethnicity/country on individual items, those items should be treated differently than the other items for the purpose of scale construction and measurement.

- This study is designed to investigate the impact of country Differential Item Functioning (DIF) on the construction of an International Mathematics Scale (IMS).
Differential Item Functioning

- Differential item functioning (DIF) happens when examinees having identical levels on the latent trait that the test was designed to measure but belonging to different groups (e.g., gender, ethnicity), have different probabilities to answer a particular item correctly (Wang, 2008).

- When a test contains items that display DIF, the test no longer measures the same latent trait for different groups of examinees and the test can no longer be considered equivalent across groups.

- As Wang (2008) indicates, real tests can never be perfect and always include DIF to some extent. In practice, as long as the magnitude of DIF is reasonably small, then the test can practically be considered as invariant.
An Example of a Country DIF-free Item

Educational Assessment Australia
An Example of an Item Displaying Country DIF:
In favour of Country D
An Example of an Item Displaying Country DIF:
In favour of Country A
An Example of an Item Displaying Gender DIF:
In favour of Country C & B
Items displaying DIF not only increase the equating error but could also be biased towards some examinees.

Looveer & Mulligan (2011) found that items used to link test forms must not display differential item functioning across year levels.

As a result, it is crucial to detect items exhibiting DIF and deal with them before performing test equating and scaling to ensure test fairness (Chu & Kamata, 2003).
Objective

- The main purpose of this study was to use the Rasch measurement model to investigate the impact of country Differential Item Functioning (DIF) on the construction of an International Mathematics Scale (IMS).
Methodology

- Instruments
- Participants
- Measurement Model
- Statistical Procedure
- Purification Approach
Instruments

- Mathematics tests (Grade 3-12) are used in this study.
- These tests are designed to measure mathematical skills in a range of contexts from the following areas:
  - number and arithmetic
  - algebra and patterns
  - measures and units
  - space and geometry
  - chance and data
A total of 20,822 students (Grade 3 to Grade 12) were randomly selected from four countries that took an International Mathematics Test in 2010.

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2246</td>
</tr>
<tr>
<td>4</td>
<td>2246</td>
</tr>
<tr>
<td>5</td>
<td>2248</td>
</tr>
<tr>
<td>6</td>
<td>2249</td>
</tr>
<tr>
<td>7</td>
<td>2204</td>
</tr>
<tr>
<td>8</td>
<td>2249</td>
</tr>
<tr>
<td>9</td>
<td>2137</td>
</tr>
<tr>
<td>10</td>
<td>2236</td>
</tr>
<tr>
<td>11</td>
<td>1956</td>
</tr>
<tr>
<td>12</td>
<td>1051</td>
</tr>
<tr>
<td>Total</td>
<td>20822</td>
</tr>
</tbody>
</table>
The Rasch simple logistic model was applied to analyse the data using RUMM 2030 software (Andrich, Lyne, Sheridan & Luo, 2005). RUMM 2030 uses the analysis of variance of the residuals to identify those items with country DIF.

Rasch analysis was carried out at two levels to examine:

1) the impact of country DIF among all items on the construction of an IMS, and
2) the impact of country DIF among link items on construction of an IMS.
There are different statistical procedures to identify differentially functioning test items (Clauser & Mazor, 1998).

The statistical test used for detecting DIF is an ANOVA of the person-item deviation residuals with person factors (e.g., country) and class intervals (e.g., group along the trait) as factors.

Once DIF was identified using an ANOVA of the person-item deviation residuals, the purification approach (Tennant & Pallant, 2007) was adopted.
Purification Approach

- In this approach, the pure items set were identified by removing items displaying DIF.

- Then, the item parameter estimates for pure items were exported to an anchor file.

- The original full set of items was then re-run while anchored by those pure items so that person estimates were based upon the measurement framework defined by the anchored items which show the minimum DIF.

- Finally, the person measures obtained from the two analyses (original and original-with-anchors) were compared.
As discussed previously, Rasch analysis was carried out at two levels to examine the impact of DIF on the scale construction:

- Level 1: The impact of country DIF among all items, and
- Level 2: The impact of country DIF among link items.

Results of the first level will be presented in this section.
Figure 1: Progression in Maths from Grade 3 to Grade 12 in Country A
Figure 2: Progression in Maths from Grade 3 to Grade 12 in Country B
Figure 3: Progression in Maths from Grade 3 to Grade 12 in Country C
Figure 4: Progression in Maths from Grade 3 to Grade 12 in Country D
Figure 5: Progression in Maths from Grade 3 to Grade 12 Before Adjustment for DIF
Figure 6: Progression in Maths from Grade 3 to Grade 12 After Adjustment for DIF
This study has applied the Rasch measurement model to investigate the impact of country Differential Item Functioning (DIF) on the construction of an International Mathematics Scale (IMS).

The results of this study show that country DIF has a significant impact on the scale construction and can change the relative position of participating countries on the latent variable.

Where data from different countries are to be pooled for constructing an international scale, it is essential to carry out DIF analysis, and if necessary, to adjust for DIF before performing test equating.
The End

Thank you for participating
Are there any questions, please?