Masking Distinct and Reliable Subscores: A Call to Assess Added Value Invariance

Joseph Rios
Educational Testing Service
Statement of Problem

• In practice, raw subscores that lack both distinctiveness and reliability are often reported

• A number of different approaches have been proposed with little success at the operational-level (e.g., augmentation procedures)

• Some researchers have concluded that without massive changes in test development, reporting subscores may not be a worthwhile endeavor
Limitations of Subscore Research

• Previous research has assumed that the validity of diagnostic information is invariant across all subgroups

• In practice, subscore added value invariance is rarely evaluated

• Sinharay and Haberman (2014) have found differential added value for ethnic and linguistic groups in operational testing contexts
Purpose of Study

The purpose of this study was to highlight how the assumption of invariance can lead to incorrect subscore added value conclusions for a large group of examinees.

When considering the total sample, what is the proportion of examinees with known added value that can be present without concluding that there is subscore added value for everyone?
METHOD
Data Generation

• Data were generated separately for two groups administered an \( n \) multiple-choice item test comprised of four subdomains

• The two groups simulated in this study differed on the degree of multidimensionality underlying the ability estimates on the four subdomains

  • Group 1 possessed a unidimensional representation of the four subdomains by having inter-subdomain correlations of .90

  • Group 2 possessed inter-subdomain correlations that ranged from weak to moderate (using Cohen’s, 1968 criteria)
Data Generation

• A correlated-traits 3-PL IRT model was used to generate item responses using the *MIRT R* package.

• Theta values were obtained from a multivariate standard normal distribution.

• Item parameters were taken from the operational administration of the 2013 MCAS.
Assessing Masking Effects

• To avoid the confounding of results, it was important to first assess the necessary conditions for subscore added value *solely* based on the generated multidimensional data.

• Next, both unidimensional and multidimensional cases were combined at different proportions to determine the degree of masking effects for cases with added value when only evaluating the total sample.

• Haberman’s (2008) method was applied to assess subscore added value.
Independent Variables

- Proportion of multidimensionality
  - 10%, 20%, 30%, 40%, 50%

- Degree of multidimensionality
  - 0.4, 0.5, 0.6

- Subdomain test length (4 total subdomains)
  - 10, 25, 50
    - Sinharay (2010) found that operational tests have subdomain test lengths that range from 10 to 70 items

- For each condition 2,500 simulees were generated and replicated 100 times
RESULTS
## Necessary Conditions for Added Value

<table>
<thead>
<tr>
<th>Test Length</th>
<th>( r )</th>
<th>( \alpha_{tot} )</th>
<th>PRMSE(_S)</th>
<th>PRMSE(_X)</th>
<th>% Subscore Added Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.40</td>
<td>.63</td>
<td>.44</td>
<td>.44</td>
<td>53%</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>.66</td>
<td>.45</td>
<td>.50</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.69</td>
<td>.45</td>
<td>.56</td>
<td>0%</td>
</tr>
<tr>
<td>25</td>
<td>.40</td>
<td>.81</td>
<td>.67</td>
<td>.47</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>.83</td>
<td>.67</td>
<td>.54</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.85</td>
<td>.67</td>
<td>.61</td>
<td>100%</td>
</tr>
<tr>
<td>50</td>
<td>.40</td>
<td>.90</td>
<td>.80</td>
<td>.50</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>.91</td>
<td>.80</td>
<td>.57</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td>.92</td>
<td>.80</td>
<td>.65</td>
<td>100%</td>
</tr>
</tbody>
</table>
## Masking Effects:
### 25 items per subdomain

<table>
<thead>
<tr>
<th>% Multidim</th>
<th>% Subscore</th>
<th>Added Value</th>
<th>% Subscore</th>
<th>Added Value</th>
<th>% Subscore</th>
<th>Added Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>30%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>40%</td>
<td>19%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>50%</td>
<td>86%</td>
<td>22%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
# Masking Effects: 50 items per subdomain

<table>
<thead>
<tr>
<th>% Multidim</th>
<th>Inter-Subdomain Correlations</th>
<th>% Subscore</th>
<th>% Subscore</th>
<th>% Subscore</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_{G1} = .90$</td>
<td>$\rho_{G1} = .90$</td>
<td>$\rho_{G1} = .90$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\rho_{G2} = .40$</td>
<td>$\rho_{G2} = .50$</td>
<td>$\rho_{G2} = .60$</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td>68%</td>
<td>23%</td>
<td>2%</td>
</tr>
<tr>
<td>30%</td>
<td></td>
<td>100%</td>
<td>96%</td>
<td>42%</td>
</tr>
<tr>
<td>40%</td>
<td></td>
<td>100%</td>
<td>100%</td>
<td>97%</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Masking Effects

Total Sample Disattenuated Inter-Subdomain Correlation

Generating Focal Group Inter-Subdomain Correlation

- 10% Focal
- 20% Focal
- 30% Focal
- 40% Focal
- 50% Focal
Multidimensional Score Profiles Masked by Correlational-Based Methods

Raw Score Difference from Mean

Score Profile

Subdomain

1 2 3 4
DISCUSSION
Summary

• Assuming invariance for all examinees can lead to incorrect conclusions for as much as 50% of the total sample

• The dichotomous framing of subscores either being valid or invalid for all examinees is based on the dependency of relying on correlations, which is a group-based index

• In doing so, one may neglect: a) possible subscore variability among subgroups or individual examinees, and/or b) differences in subtest difficulty (Raymond & Feinberg, 2015)
Implications

• Evaluate subscore added value invariance for identifiable and protected minority groups

• This approach assumes subgroup homogeneity and is limited in two ways
  
  • Depending on how the subgroup is split or classified, restriction of range may be a concern leading to attenuated correlations and reliability estimates
  
  • As with any group-based method, it may ignore subscore variability among individuals within the subgroup
Future Research Directions

• An alternative approach could be to evaluate subscores based at the individual-level
  • Raymond and Feinberg (2015)

• Need to better understand how stakeholders are using subscores to improve instruction and learning
  • Incorporate more research from the feedback intervention theory literature
Thank You

jrios@ets.org