DIF AND SGPS: IMPLEMENTING THE POPHAM-RYAN DESIGN

Stephen C. Court
Educational Research and Evaluation, LLC

A Presentation at the First International Conference on Instructional Sensitivity

Achievement and Assessment Institute
University of Kansas, Lawrence, Kansas, November 13-15, 2013
Popham & Ryan Design

- Utilizes existing data
- Applies statistical procedures that are practical for most settings

**Item x Grouping w/Layer**

**Items:** 2013 Grade 5 and Grade 8 Reading and Math

**Teacher Grouping:** Median SGP (50-50)

- based on 2011 and 2012 test results

**Layer:** 2013 Proficient/Not Proficient
**Purpose is not:** To use the item to validate the accuracy of the grouping method.

**Purpose:** To appraise the instructional sensitivity of the item. Therefore, accept the grouping method as accurate...for the sake of argument.
## Appraisal Results

<table>
<thead>
<tr>
<th>Content</th>
<th>$N_{\text{students}}$</th>
<th>$N_{\text{items}}$</th>
<th>Insensitive</th>
<th>Marginally Sensitive</th>
<th>Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>1268</td>
<td>69</td>
<td>66%</td>
<td>25%</td>
<td>9%</td>
</tr>
<tr>
<td>Reading</td>
<td>4514</td>
<td>82</td>
<td>82%</td>
<td>18%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Grade 8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>988</td>
<td>84</td>
<td>47%</td>
<td>38%</td>
<td>15%</td>
</tr>
<tr>
<td>Reading</td>
<td>5297</td>
<td>81</td>
<td>88%</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td>71%</td>
<td>23%</td>
<td>6%</td>
</tr>
</tbody>
</table>
Classification based on significance level:

- Sensitive: < .001
- Marginally sensitive: .001 to .05
- Insensitive: > .05

- Typical MH Estimate range for Insensitive: 1.000
- Typical MH Estimate range for Marginally Sensitive: 1.250
- Typical MH Estimate threshold for Sensitive: > 1.500
Corresponding ROC analysis associates sig. levels of .000 with AUC values of just .560.

Yet, AUC values typically interpreted as:

- .90-1.0 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)
Sig. levels derived from N-based standard errors may not serve well as basis of interpretation. Alternative schemes for interpreting MH Estimates:

<table>
<thead>
<tr>
<th>MH Range</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 2 (inc. values &lt; 1)</td>
<td>Fail</td>
</tr>
<tr>
<td>2 to 5</td>
<td>Poor</td>
</tr>
<tr>
<td>5 to 10</td>
<td>Good</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MH Range</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 5</td>
<td>Insensitive</td>
</tr>
<tr>
<td>5 to 10</td>
<td>Marginally Sensitive</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>Sensitive</td>
</tr>
</tbody>
</table>
Comparisons

Teacher Grouping

- Baseline: Random Equivalent Halves

  0% Sensitive; 100% Insensitive

No discrimination; complete overlap
Two consecutive years of class proficiency above 80%.

Extreme discrimination between groups:

<table>
<thead>
<tr>
<th>Grouping Method</th>
<th>Between group Mean difference</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGP</td>
<td>.004</td>
<td>-.01</td>
<td>.02</td>
</tr>
<tr>
<td>Proficiency</td>
<td>.203</td>
<td>.04</td>
<td>.51</td>
</tr>
</tbody>
</table>

The mean difference between the two Proficiency groups was more than 50 times greater than the mean difference between the two SGP-based teacher groups.
The proficiency-based grouping method tended to be more sensitive to instructional sensitivity.

For grade 5 reading:
In practice, we do **not** design tests and create items for extremely separated groups.

We design tests and create items for everyone – all schools, all teachers, and all students.

The fact that the method works for extreme groups provides evidence that the SGP-DIF method proposed by Popham and Ryan yields a basis for valid appraisals of item sensitivity:
Except under the most extreme conditions, the items are not sufficiently sensitive to instruction – i.e., to contrasts of more effective vs. less effective teachers.

In turn, they should not contribute to test scores that serve as the basis for high-stakes decisions regarding:

- school effectiveness
- instructional quality
- teacher competence.
We would not render summative judgment of a teacher’s competence on the basis of a just one set of test scores.

Therefore, nor should we render summative judgment on an item’s sensitivity to instruction on the basis of one appraisal approach.

Instead, use five methods – e.g., MH, ROC, LR, MI, and BSEM:

1. For each method, classify items as $0 = \text{Insensitive}; 1 = \text{Marginally Sensitive};$ and $2 = \text{Sensitive}.$

2. Summing the classifications $\rightarrow$ a sensitivity index from 0 to 10.

3. Review items with sum $< X,$ where $X$ is set judgmentally on the basis of test purpose and stakes attached.

   For high-stakes, such as teacher evaluation, $X$ might be set at 7 or 8.
Examine actual items for contextual characteristics that explain why the items function as they do under

- Extreme conditions
- Real-world conditions

Examine not just p-values, biserials, reliability coefficients, sensitivity statistics, etc. but also actual item and stimulus format and content – for example:

- **Stems:** question vs. completion
- **Options:** short vs. long
- **Alignment:** Webb, Porter, etc.
- **Cognitive load:** DOK, Bloom, etc.