Similarities and differences between groups of countries concerning relative weaknesses and strengths

GA meeting, Wien October 2014

Svein Lie
University of Oslo
The role of achievement test items

- A test consists of numerous test items that together represent a good measurement of the actual competency.

- This collection of items must ensure high validity as well as reliability:
  - Validity: good coverage of the decided framework concerning types of competencies, item formats, difficulties, etc
  - Reliability: ensuring low measurement errors, giving good estimates of their magnitude

- A high quality test score can then be calculated by combining data from many items into a score
  - How to do that depend on the detailed measurement model
A test item perspective

- From response to item score:
  - for MC items: *which* alternative chosen
  - for constructed (open) response items: *what* was written, type of correct or wrong response
    - For many items reports include this information
    - Two-digit scoring rubrics

- From item scores to scale score
  - Only "sum" of p-values "counts"
  - Valuable to investigate also the "90 % lost" information
  - Not only random noise!!!
Looking for interesting secondary information from test items

Two examples:

- Investigating student conceptual understanding by investigating responses to one or a few items
  - A number of such studies has been carried out at EKVA

- Looking for meaningful patterns in the item residuals (remains when the score aspect has been taken out)
  - An in-depth study of relative strengths and weaknesses for countries
Norway in the “Nordic river” (excl. Finland) p-values items in TIMSS 1995
Calculating p-value residuals

- How much better or worse is a national p-value than expected, considering
  - the international difficulty of the item
  - and the overall score for the country?
Looking for patterns in the residual p-value matrix

- Possible patterns of countries
  - pairs or groups of countries showing similar strengths and weaknesses

- or of items
  - groups of items that tend to discriminate similarly between countries

- Tools: correlations, cluster analysis etc.
In practice

- Very simple procedure
  - Start with a p-value matrix: item by country
  - Average p-values calculated both for countries and for items
  - Subtract actual item average as well as country average to get country/item residuals
  - How much better or worse did a particular country score than expected from the country’s overall score and the item’s overall difficulty?
### Table 1: Item Scores

<table>
<thead>
<tr>
<th>Countries</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>45</td>
<td>60</td>
<td>32</td>
<td>45.7</td>
</tr>
<tr>
<td>Item 2</td>
<td>57</td>
<td>80</td>
<td>53</td>
<td>63.3</td>
</tr>
<tr>
<td>Item 3</td>
<td>38</td>
<td>65</td>
<td>38</td>
<td>47.0</td>
</tr>
<tr>
<td>Mean</td>
<td>46.7</td>
<td>68.3</td>
<td>41.0</td>
<td>52.0</td>
</tr>
</tbody>
</table>

### Table 2: Item Differences

<table>
<thead>
<tr>
<th>Countries</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>-1.7</td>
<td>-8.3</td>
<td>-9.0</td>
<td>-6.3</td>
</tr>
<tr>
<td>Item 2</td>
<td>10.3</td>
<td>11.7</td>
<td>12.0</td>
<td>11.3</td>
</tr>
<tr>
<td>Item 3</td>
<td>-8.7</td>
<td>-3.3</td>
<td>-3.0</td>
<td>-5.0</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### Table 3: Item Normalization

<table>
<thead>
<tr>
<th>Countries</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>4.7</td>
<td>-2.0</td>
<td>-2.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Item 2</td>
<td>-1.0</td>
<td>0.3</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Item 3</td>
<td>-3.7</td>
<td>1.7</td>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Correlations of p-value residuals between Norway and all other countries. Mathematics. TIMSS 1995

<table>
<thead>
<tr>
<th>TOP 10</th>
<th>BOTTOM 10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sweden</strong></td>
<td><strong>Kuwait</strong></td>
</tr>
<tr>
<td>.68</td>
<td>-.25</td>
</tr>
<tr>
<td><strong>Iceland</strong></td>
<td><strong>Israel</strong></td>
</tr>
<tr>
<td>.55</td>
<td>-.25</td>
</tr>
<tr>
<td><strong>Denmark</strong></td>
<td><strong>Singapore</strong></td>
</tr>
<tr>
<td>.47</td>
<td>-.30</td>
</tr>
<tr>
<td><strong>Germany</strong></td>
<td><strong>Thailand</strong></td>
</tr>
<tr>
<td>.40</td>
<td>-.34</td>
</tr>
<tr>
<td><strong>Switzerland</strong></td>
<td><strong>Korea</strong></td>
</tr>
<tr>
<td>.37</td>
<td>-.34</td>
</tr>
<tr>
<td><strong>Scotland</strong></td>
<td><strong>Russia</strong></td>
</tr>
<tr>
<td>.36</td>
<td>-.35</td>
</tr>
<tr>
<td><strong>New Zealand</strong></td>
<td><strong>Hong Kong</strong></td>
</tr>
<tr>
<td>.36</td>
<td>-.37</td>
</tr>
<tr>
<td><strong>England</strong></td>
<td><strong>Bulgaria</strong></td>
</tr>
<tr>
<td>.35</td>
<td>-.38</td>
</tr>
<tr>
<td><strong>Netherlands</strong></td>
<td><strong>Iran</strong></td>
</tr>
<tr>
<td>.30</td>
<td>-.43</td>
</tr>
<tr>
<td><strong>Australia</strong></td>
<td><strong>Romania</strong></td>
</tr>
<tr>
<td>.29</td>
<td>-.44</td>
</tr>
</tbody>
</table>
Clustering with labels and reliability
TIMSS Science 1995

- **East Asia (0.52)**: Hong Kong, Japan, Korea, Singapore, Thailand
- **East Europe (0.68)**: Bulgaria, Czech rep, Hungary, Latvia, Lithuania, Romania, Russia, Slovac rep, Slovenia
- **English-speaking (0.88)**: Australia, Canada, England, Ireland, New Zealand, Scotland, USA
- **North Europe (0.73)**: Denmark, Iceland, Norway, Sweden, Belgium (FI), Netherlands, Switzerland
- **South Europe (0.44)**: Cyprus, Greece, Portugal, Spain
Cluster analysis:

- A method to cluster variables by agglomeration of cases into ever larger groups based on similarities

- Various measures of similarity, e.g.
  - Correlations (most common), «Distance» etc

- Various rules for combining groups, e.g.
  - Internal cohesion, External isolation
PISA science 2006

- Including test items as well as student attitudes towards science
Groups and countries 2006

- East Central Europe (former communist)
- English-speaking
- French/Dutch
- German-speaking
- Islamic
- Latin America
- Latin Europe
- Nordic
- North-East Asia
Why, and So what?

- Actual clustering «mechanisms»: language, geography, political and cultural history, religion, curricular influences etc

- Such groupings of countries invites to focus on:
  - Identifying what constitutes the particularity for each group, and looking for influences whithin and between groups
  - Encouriging regional (groupwise) study reports to better interpretation of national test results concerning important challenges
EKVA

Faculty of Education, University of Oslo

- Norwegian center for administering and running tests and other quantitative studies
  - Internationally: IEA studies, PISA etc
  - National assessments
  - Master- and PhD quantitative projects in science/math/reading education

- International studies:
  - Main task: administrating, running, recording according to given procedures
  - National reporting focusing on Norwegian results/standing in an international context
  - Focusing on comparison with nbour countries
  - Also focusing on trends and links between studies
Some references

Kjærnsli & Lie:

- *International Journal of Science Education*, 2011/33, p 121-144
- *IEA Research Conference, Cyprus 2005*