Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling

Steve Holmes, Qingping He and Michelle Meadows
The context

- GCSEs taken in England by students aged 16 are currently being reformed with respect to content and the associated assessment.
- Examinations in the same subject areas are provided by several exam boards.
- There is no pre-testing or equating between exams due to security considerations and their high-stakes nature.
- Small differences in difficulty can be dealt with by adjusting grade boundaries. Large differences may wash-back on teaching and learning.
- A need arose for Ofqual to evaluate reformed GCSE Maths sample assessment difficulties.
Aim of study

To explore the potential of using comparative judgement and Rasch modelling to investigate the relative difficulty between examinations
Method

- Items in six mathematics question papers designed by three exam boards (Boards A, B and C) for 16-year olds in England were judged in paired comparison by experts. Three of the papers were for the more able students (Higher Tier) and the other three for the less able students (Foundation Tier).

- The Rasch model for dichotomous items was applied to the paired comparison data to establish the scale of expected difficulty.

- The six papers were also taken by 2933 students using an equivalent-groups design for each tier, allowing the difficulties of the items to be compared and placed on the same measurement scale using the Partial Credit Model (PCM).

- The actual item difficulties derived from the test data using the Partial Credit Model were compared with the expected item difficulties derived from the comparative judgement data and the Rasch model to validate the comparative judgement approach.
The online comparative judgement process

These prisms are named after the shape of their end face.

(a) Complete this table.

<table>
<thead>
<tr>
<th>Shape of end face</th>
<th>Number of faces</th>
<th>Number of edges</th>
<th>Number of vertices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle (3 sides)</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Rectangle (4 sides)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentagon (5 sides)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hexagon (6 sides)</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

(b) How many edges and vertices does a prism with a 100-sided end face have?

Edges
Vertices

(c) Write down a formula connecting the number of faces $F$ of a prism and the number of sides of its end face $n$.

(e) 

Calculator allowed.

Person A is in a class of 28 students, 3 of whom are left-handed. There are 1,250 students in the school.

(a) Use this information to estimate how many students in the school are left-handed.

(b) Is your solution to (a) likely to be an overestimate or an underestimate? Explain your reasoning.

(c) Person B is at a different school. Person B is in a class of 25 students, 6 of whom are left-handed.

Person B says to Person A, "In our two classes, there are 54 students, 9 of whom are left-handed. We can use this bigger sample to improve the estimate."

What assumption has Person B made? Explain whether you think that Person B’s argument is correct.
Distribution of expected item difficulties derived using comparative judgement and the Rasch model
Test characteristic curves (TCCs) derived using comparative judgement data and Rasch model for dichotomous items
Question paper statistics based on test data using classical test theory

<table>
<thead>
<tr>
<th></th>
<th>Board A</th>
<th></th>
<th>Board B</th>
<th></th>
<th>Board C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>F</td>
<td>H</td>
</tr>
<tr>
<td>Number of students</td>
<td>325</td>
<td>618</td>
<td>326</td>
<td>648</td>
<td>353</td>
<td>627</td>
</tr>
<tr>
<td>Number of items</td>
<td>44</td>
<td>37</td>
<td>48</td>
<td>36</td>
<td>37</td>
<td>28</td>
</tr>
<tr>
<td>Scaled maximum available mark</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Mean scaled score</td>
<td>40.16</td>
<td>33.91</td>
<td>24.44</td>
<td>27.98</td>
<td>23.62</td>
<td>18.04</td>
</tr>
<tr>
<td>Standard deviation of scaled score</td>
<td>18.53</td>
<td>16.89</td>
<td>15.82</td>
<td>16.96</td>
<td>13.14</td>
<td>13.01</td>
</tr>
<tr>
<td>Cronbach's alpha</td>
<td>0.90</td>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>McDonald’s omega_t</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Distributions of scores (scaled to 100)

**Foundation Tier**

- Frequency vs. Test score (scaled)
  - Board A
  - Board B
  - Board C

**Higher Tier**

- Frequency vs. Test score (scaled)
  - Board A
  - Board B
  - Board C
Ability and difficulty distributions based on PCM analysis of test data
Distribution of observed item difficulties derived using test data and PCM
Test characteristic curves (TCCs) based on PCM analysis of test data
Comparison of Test characteristic curves (TCCs)

Test Data

TCC based on actual test data, F tier

TCC based on actual test data, H tier

CJ Data

TCC based comparative judgement, F tier

TCC based comparative judgement, H tier
Comparison of expected item difficulties derived using comparative judgement data and observed difficulties derived using test data

Correlation of 0.66 and disattenuated correlation of 0.76
Concluding remarks

- The expected item difficulties derived using the comparative judgement data and Rasch model and the actual item difficulties derived using the test data and PCM were reasonably strongly correlated.
- Comparative judgement appears to be an effective way to investigate the comparability of difficulty between examinations.
- It could be used as a proxy for pretesting high-stakes tests in situations where pretesting is not feasible.
- There may be scope for refining the judging criteria.
- We have indeed used the comparative judgement approach alone to assess difficulty in a few assessments.
Thank you