Balancing statistical evidence with expert judgement when aligning tests to the CEFR

Professor Anthony Green
CRELLA University of Bedfordshire

Ed Hackett
Head of Test Development
Oxford University Press

Date: 1 June 2014
EALTA – Warwick University
Potential sources of evidence

- Item writers
- Expert judges
- Teachers
- Test takers
- Statistics
Case Study
The Oxford Test of English B

www.oxfordtestofenglish.com
What is the Oxford Test of English B?

• General proficiency test
• CEFR A2, B1, and B2
• Part of OUP’s CEFR aligned assessment and course provision

Placement test
Learning materials
What is the Oxford Test of English B?
General Proficiency Test CEFR A2-B2

• Institutional use
• 100% online
• On-demand
• Algorithm-driven
• Flexible format
## The Oxford Test of English B

### Four modules

<table>
<thead>
<tr>
<th>Module</th>
<th>CEFR</th>
<th>No. Parts</th>
<th>No. Tasks</th>
<th>No. Items</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>A2, B1, B2</td>
<td>4</td>
<td>9</td>
<td>22</td>
<td>Approx. 30 mins</td>
</tr>
<tr>
<td>Listening</td>
<td>A2, B1, B2</td>
<td>4</td>
<td>12</td>
<td>20</td>
<td>Approx. 30 mins</td>
</tr>
<tr>
<td>Speaking</td>
<td>A2 - B2</td>
<td>4</td>
<td>6</td>
<td>15</td>
<td>Approx. 15 mins</td>
</tr>
<tr>
<td>Writing</td>
<td>A2 - B2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Approx. 45 mins</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>29</strong></td>
<td></td>
<td><strong>59</strong></td>
<td><strong>Approx. 2 hours</strong></td>
</tr>
</tbody>
</table>
Results
Linked to CEFR

Overall CEFR level

CEFR level for each skill

Online verification
Linking Options

Sources of evidence for aligning tests
Starting points
Oxford Test of English B (OTE-B) and the CEFR

• OTE-B based on the CEFR: embedded in test development process
  – Items designed to target a CEFR level
• Need for coherence within the Oxford product range: common interpretation of levels across tests
  – Oxford Online Placement Test (OOPT) previously linked to CEFR
• International test: broad coverage for diverse test taking population
  – Linking to take account of diversity
Linking the OTE-B to the CEFR

Stages in development and linking evidence

• Test specifications
• Trial materials
• Item writer guidelines
• Item writing process (Writing, Editing, Vetting)
Four perspectives on the OTE-B: CEFR relationship

Do all point to the same conclusion?
A priori evidence from test material
Evidence from item production

- Item writers familiarised with CEFR
- *Item Writer Guidelines* specify language and functions at each level
- OTE-B items developed (written, edited, revised) to operationalise CEFR descriptors
- Materials compiled to target a CEFR level: A2, B1, or B2
A priori evidence from test material
The panel of expert judges

- Reading and Listening tasks

- Expert panel
  - Recruited a group of 12 ‘Expert’ Judges from 3 backgrounds:
    - EFL Testing Academics
    - EFL Teachers
    - EFL Materials Writers

- CEFR Training Exercise
Test material and test taker performance

1. Judge difficulty
2. Give test

Item difficulty NOT known to judges

Threshold range

B1
B1
A2
B1
A2
B1
A2

0.44 0.42 0.39 0.37 0.34 0.33 0.31 0.28 0.24

1. Give test
2. Judge cut point

Item difficulty known to judges
Evidence from anchor items
Linking with the Oxford perspective

- Oxford Online Placement Test previously aligned to CEFR via student Can Do and teacher ratings
- OOPT in operation for 5 years: stable difficulty, stakeholders satisfied with interpretation of CEFR
- Selected OOPT items of known difficulty seeded into the OTE-B as anchor items
- Pre-test administered to a representative sample across mix of L1s
- Cut score for OTE-B can initially be set according to the levels determined for OOPT through Rasch scaling
Evidence based on learner performance
Person-based: contrasting groups

- Test takers of known ability
- Compare performance of B1 level learners with A2 level learners
- Cut score is located at the intersection between the two groups

Threshold range

<table>
<thead>
<tr>
<th>A2</th>
<th>A2</th>
<th>A2</th>
<th>B1</th>
<th>A2</th>
<th>B1</th>
<th>A2</th>
<th>B1</th>
<th>A2</th>
</tr>
</thead>
</table>

Judges are teachers
1. Judge student ability
2. Give test
Balancing the evidence

Where should we locate the cut points?
The item production perspective:
Difficulty of items written to target CEFR levels

• OTE-B items written to target a CEFR level: A2, B1 or B2
• Items approved and selected as representing CEFR level
The item production perspective:
Difficulty of items written to target CEFR levels

A2
B1
B2

B1: 43.5
B2: 63.9
The expert judge perspective: Method

- 12 expert judges rated items across three CEFR levels: A2, B1, and B2
- Each item rated at low, mid, high within each target CEFR level
- Analysis of expert judge performance using multi-facet Rasch: one misfitting judge (Outfit MnSq > 1.4)
- Judgements were scaled and correlated with difficulty values obtained through pretesting
Correlation of expert judgements with pre-test values

Correlation = 0.744

\[ y = 0.0657x + 24.02 \]

Two extreme outliers eliminated

B1: 49.9
B2: 61.3
Alignment Issues
Reading and Listening scale

Three sources of evidence
Three conclusions
Which one should we trust?
Balancing Evidence – Triangulation

Test Scores

Anchored pre-test scale

Experts’ ratings of tasks

Teachers’ ratings of learners
Pilot Stage

- Pilot full test with 300+ test takers
- Teacher ratings
- Correlation of teacher ratings with test performance data
- Triangulation: Anchored Pretest, Expert Judgement and Teacher ratings of learners

- Which scales most closely align?
- Are results consistent across skills?
- Is there any evidence of bias?
Summary

- Different sources of evidence can provide different answers

- Need to evaluate and balance three perspectives:
  - People interpret the CEFR to arrive at cut points
  - Test results provide an order of difficulty
  - Piloting grounds the scale in the learning environment
Thank you

For more information:

www.oxfordtestofenglish.com