How do we balance statistical evidence with expert judgement when aligning tests to the CEFR?

Professor Anthony Green
CRELLA University of Bedfordshire

Colin Finnerty
Senior Assessment Manager
Oxford University Press

Date: 11 April 2014
ALTE - Paris
Potential sources of evidence

- Item writers
- Expert judges
- Teachers
- Test takers
- Statistics

\[(a + b)^2 = a^2 + 2ab + b^2\]
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
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>TOTAL</td>
<td></td>
<td>2</td>
<td>29</td>
<td>59</td>
<td>Approx. 2 hours</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?

Learner performance

Anchor items

Test material

Test material plus performance data
A priori evidence from test material
Evidence from item writers

- Item writers familiarised with CEFR
- OTE-B items developed to operationalise CEFR descriptors
- *Item Writer Guidelines* specify language and functions at each level
- Materials written to target a CEFR level: A2, B1 or B2
A priori evidence from test material

The panel of expert judges

- Selected range of Reading and Listening tasks with known difficulties
- Recruited a group of 12 ‘Expert’ Judges from 3 backgrounds:
  - EFL Testing Academics
  - EFL Teachers
  - EFL Materials Writers
- CEFR Training Exercise
A priori evidence from test material
Evidence from expert judges

- Expert judges rate difficulty of test material
  
  To which performance level description (i.e. CEFR level or category) are the knowledge, skills and cognitive processes required to respond successfully to this item most closely matched?

- Angoff-type methods: Cut score for B1 = sum of items that CEFR B1 test taker is judged to be able to answer
A *posteriori* evidence from test material and test taker performance

Bookmark type methods: IRT methods

• After empirical item difficulty becomes known (post-pretest anchoring)
• Items ordered according to difficulty
• Standard setting panel judges for each item whether the probability of a ‘borderline person’ giving a correct answer is at or above a set ‘probability threshold’ (e.g. \( \frac{2}{3} \))
• Test based: cut score set at point in the test at which experts judge probability of a correct response falls below the threshold
• Score based: cut score set at point in the threshold range (e.g. where items judged at A2 and at B1 overlap).
Test material and test taker performance

1. Judge difficulty
2. Give test

Item difficulty **NOT** known to judges

Threshold range

<table>
<thead>
<tr>
<th></th>
<th>A2</th>
<th>B1</th>
<th>A2</th>
<th>B1</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.28</td>
<td>0.24</td>
<td>0.31</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.39</td>
<td>0.39</td>
<td>0.42</td>
<td>0.44</td>
</tr>
</tbody>
</table>

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

Judges are teachers
1. Judge student ability
2. Give test

Threshold range
Balancing the evidence

Where should we locate the cut points?
Correlation between expert judgement and anchored values

Method

- 12 expert judges rated 27 common items across three CEFR levels A2, B1, and B2
- Each item rated at low, mid, high within the CEFR level
- Analysis of expert judges
- 7 ‘best’ expert judges selected for correlation exercise with Pretest anchored values
Correlation of expert judgements with pre-test values

Correlation = 0.76
Alignment Issues
Reading and Listening scale

• Expert ratings suggest CEFR cut points for A2, B1 and B2 could be revised downward by \( \frac{1}{4} \) to \( \frac{1}{2} \) a CEFR level.

• Which scale do we trust?
Balancing Evidence – Triangulation

- Anchored pre-test scale
- Experts’ ratings of tasks
- Teachers’ ratings of learners

Test Scores
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