Setting multiple CEFR cut scores for assessments intended for young learners

Patricia A. Baron & Spiros Papageorgiou
Educational Testing Service

11th EALTA Conference, University of Warwick, UK
May 29-June 1, 2014
Rationale

• Mapping language test scores to the CEFR typically aims to help score users interpret students’ level of language proficiency
  – “the six main levels of the CEFR have become a common currency in language education” (Alderson, 2007, p. 660)
  – reference to its scales can “add meaning to the scores” (Kane, 2012, p.8)
• Mapping of test scores on the CEFR for an EFL assessment for young learners (Hasselgreen, 2005, 2012)
TOEFL Primary

• Purpose: measure reading, listening and speaking skills of young EFL learners ages 8+
  – Paper-delivered Reading and Listening (36 items, two steps)
  – Computer-delivered Speaking (7 tasks)

• Test Development
  – Analysis of curriculum of young learners, 9 countries
  – Identified communication goals for this age group
  – Prototyped tasks, pilot tested items

• Intended uses
  – Assess general language proficiency
  – Support teaching and learning
  – Not high stakes
TOEFL Primary Mapping

• Standard setting study
  – Minimum scores (cut scores) needed to reach specific performance levels

• Specific levels
  – CEFR levels suggested by Test Developers and ELL Researchers based on assessment design and purpose
  – Reading and Listening: A1, A2 and B1
  – Speaking: A1, A2, B1 and B2
Limitations of the CEFR for assessing young learners

- Descriptors not appropriate for younger learners when assessing higher language ability levels and competencies other than linguistic ones (Hasselgreen, 2005; Papageorgiou, 2010)
- Need for modified descriptors to make them more comprehensible and better adapted to the experiences of young learners (Hasselgreen, 2012; Schneider & Lenz 2000)
Research Questions

• To what extent do panelists refer to CEFR-based descriptors, modified for young learners, during the process of setting cut scores?

• Is the decision-making consistency satisfactory when the modified descriptors are used?
Standard Setting Panelists

• 18 educators from 15 countries
• Teachers of English as a second or foreign language (10)
• Administrative positions (8)
• Experience teaching young learners
  – less than five years (2)
  – between five and ten years (7)
  – more than ten years of experience (9)
Standard Setting Methods

• Yes/No Angoff for Reading and Listening tests
  – Would a student at each CEFR level know the correct answer?

• Performance Profile for Speaking
  – What is the CEFR level of a test taker based on the students’ scored responses?
Process

• Two preparatory activities
  – Pre-work assignment to review CEFR levels, note what differentiates levels
  – TOEFL Primary administered the day before the standard setting meeting

• During workshop
  – Define Just Qualified or Borderline student
  – Training, practice, two rounds of judgments
Process

• Modified CEFR descriptors consulted for:
  – Preparatory CEFR activity
  – Definition of Just Qualified Candidate
• Recordings of standard setting panel discussions
<table>
<thead>
<tr>
<th>Modified Descriptors</th>
<th>Original Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LISTENING B1</strong></td>
<td></td>
</tr>
<tr>
<td>Can understand straightforward factual information about common everyday and <strong>school-related topics</strong>, identifying both general messages and specific details, provided speech is clearly articulated.</td>
<td>Can understand straightforward factual information about common everyday or <strong>job-related topics</strong>, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent.</td>
</tr>
<tr>
<td><strong>READING B1</strong></td>
<td></td>
</tr>
<tr>
<td>Can understand straightforward texts, <strong>controlled for language construction and accessibility</strong>, on familiar topics.</td>
<td>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</td>
</tr>
</tbody>
</table>
Sample TOEFL Primary

• Speaking
  – *Listen to one sample response*
  – *Student got a 5 out of 5 on the scoring rubric*
Reported use of the modified descriptors

- Reading – Describing borderline A2
  - *It is about inferring, A1 is recognize, A2 is understand the meaning of high frequency words, in which topics, in which domain? We need to include descriptions or something. Is it matching? ...*

- *panel was struggling... and then:*
Actually, here it is, in the modified CEFR descriptors: it categorizes ... the first one is simple text, then simple instruction, the other one is personal correspondence, so we can also include this one, the personal, like email....... Like asking for information, complaining, thinking, can identify overall message, can identify overall idea, in short message, or short text. Message regarding common everyday life.
Reported use of the modified descriptors

• Reading – Describing borderline A2
  – need visual support? *I think they can do more than that.*.. *They can follow instructions, they can do pretty much those things in the computer game, they have to read the instructions to get to the next level.*
Reported use of the modified descriptors

• Listening – Describing borderline A2
  – What about daily activities?....
  – So if you can give some examples. They mention shared international words [in the modified descriptor], so common words. Maybe movies, print, delete...
  – We need to include descriptions or something.
Reported use of the modified descriptors

• Listening - continued
  – So why not say short, simple texts – about everyday life and school life... Wait, it is not everyday life, this is students, this is high frequency words from home and school life......familiar topics such as animals, family, sports, daily routine, food, entertainment
# Decision making consistency

### TOEFL Primary Listening

<table>
<thead>
<tr>
<th>Cut scores</th>
<th>SEJ</th>
<th>SEJ/SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.5</td>
<td>0.18</td>
</tr>
<tr>
<td>A2</td>
<td>1.4</td>
<td>0.51</td>
</tr>
<tr>
<td>B1</td>
<td>0.8</td>
<td>0.29</td>
</tr>
</tbody>
</table>

**SEM = 2.8**

SEJ – standard error of judgment; SEM – standard error of measurement

¹Based on field test item calibration (n>1200 per item). Selected 57 items to represent test blueprint.
## Decision making consistency

TOEFL Primary Listening

<table>
<thead>
<tr>
<th>Cut scores</th>
<th>SEJ</th>
<th>SEJ/SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.5</td>
<td>0.18</td>
</tr>
<tr>
<td>A2</td>
<td>1.4</td>
<td>0.51</td>
</tr>
<tr>
<td>B1</td>
<td>0.8</td>
<td>0.29</td>
</tr>
</tbody>
</table>

SEM = 2.8

---

1 Based on field test item calibration (n>1200 per item). Selected 57 items to represent test blueprint.
Decision making consistency  
TOEFL Primary Reading

<table>
<thead>
<tr>
<th>Cut scores</th>
<th>SEJ</th>
<th>SEJ/SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.9</td>
<td>0.34</td>
</tr>
<tr>
<td>A2</td>
<td>1.1</td>
<td>0.41</td>
</tr>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.15</td>
</tr>
</tbody>
</table>

SEM = 2.7

1 Based on field test item calibration (n>1200 per item). Selected 57 items to represent test blueprint.
## Decision making consistency

TOEFL Primary Speaking

<table>
<thead>
<tr>
<th>Cut scores</th>
<th>SEJ</th>
<th>SEJ/SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.2</td>
<td>0.09</td>
</tr>
<tr>
<td>A2</td>
<td>0.4</td>
<td>0.18</td>
</tr>
<tr>
<td>B1</td>
<td>0.4</td>
<td>0.18</td>
</tr>
<tr>
<td>B2</td>
<td>0.3</td>
<td>0.14</td>
</tr>
</tbody>
</table>

SEM = 2.2
Discussion

• Panelists referred to modified CEFR-based descriptors when defining performance at the cut score.
• The estimates of judgment consistency for the cut scores in our study are in the range considered acceptable, compared to SEM.
• Facilitators of similar standard setting panels might want to prepare modified CEFR descriptors to help panelists with their task.
Implications

• By doing what the literature says, i.e. modifying the CEFR descriptors to make them more relevant to the experiences of young learners, have we actually departed too much from the central meaning of the CEFR levels?
Thank you

- Resource for more information
- http://www.ets.org/toefl_primary/content