The Challenge of Relating National Grading in Examinations to the CEFR

EALTA 30.5.2014
Taina Juurakko-Paavola
taina.juurakko-paavola@hamk.fi
Matriculation Examination Board, Finland
HAMK University of Applied Sciences
Outline

1. Background
2. About the standard setting process
3. Reliability and validity of the results
4. Results
5. Conclusion
I Background

- The Matriculation Examination Test is the only high-stakes test in Finland (grade 12, age 18-19)

- the purpose of the examination is to discover whether pupils have assimilated the knowledge and skills required by the curriculum for the upper secondary school and whether they have reached an adequate level of maturity in line with that school's goals. Passing the Matriculation Examination entitles the candidate to continue his or her studies at university. (Ylioppilastutkintolautakunta)

- Almost all upper secondary school students take part in the English test, about 39 000 participants.
Background

- For language studies, the goal definition contains target levels adapted from the CEFR (11 levels: A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, **B2.1 target level in English**, B2.2, C1.1)
- The test comprises 4 parts: Listening Comprehension, Reading Comprehension, Vocabulary and Grammar, Writing
- The grades are given applying a modified normal distribution: norm-referencing, not criterion-referencing
About the Standard Setting Process

Aims of the Standard Setting Process

1) To enhance and ensure the quality of the tests
2) To enhance the comparability of the results to the levels of the CEFR and to the levels of the Finnish version of the CEFR
3) To give feedback information for the new versions of the curricula under construction
Stages of the Standard Setting Process

- procedures recommended in the Manual (CoE 2009) were applied selectively as the raters were quite familiar with the CEFR

1. Planning and co-ordination
2. Preparation of the familiarization and training materials for the linkage process
3. Familiarization and standard setting training for the item writers and raters
4. Descriptor sorting and discussion of ratings
5. Responding to the tests
6. Rating the tasks, items and a sample of writing scripts
7. Analyzing the results and reporting
A standard setting method, developed and successfully used by Kaftandjieva (2010) in Finland, Catalonia and Norway was used to set cut scores. Cumulative compound method, which does not require the use of IRT.
3. Evidence for Reliability and Validity
**Table 1. Correlations between descriptor ratings and their original CEFR levels**

<table>
<thead>
<tr>
<th>Rater</th>
<th>Listening Comprehension</th>
<th>Reading Comprehension</th>
<th>Writing</th>
<th>Grammar</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.980</td>
<td>.983</td>
<td>.970</td>
<td>.960</td>
<td>.983</td>
</tr>
<tr>
<td>2</td>
<td>.964</td>
<td>.983</td>
<td>.979</td>
<td>.976</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>.929</td>
<td>.923</td>
<td>.965</td>
<td>.942</td>
<td>.940</td>
</tr>
<tr>
<td>4</td>
<td>.864</td>
<td>.912</td>
<td>.990</td>
<td>.960</td>
<td>.973</td>
</tr>
<tr>
<td>5</td>
<td>.922</td>
<td>.869</td>
<td>.957</td>
<td>.899</td>
<td>.969</td>
</tr>
<tr>
<td>6</td>
<td>.941</td>
<td>.895</td>
<td>.905</td>
<td>.946</td>
<td>.909</td>
</tr>
<tr>
<td>7</td>
<td>.936</td>
<td>.941</td>
<td>.974</td>
<td>1.00</td>
<td>.982</td>
</tr>
<tr>
<td>8</td>
<td>.944</td>
<td>.842</td>
<td>.990</td>
<td>1.00</td>
<td>.812</td>
</tr>
<tr>
<td>9</td>
<td>.957</td>
<td>.852</td>
<td>.928</td>
<td>.875</td>
<td>.983</td>
</tr>
</tbody>
</table>
Table 2. Reliability of the ratings: Cronbach’s alfa – α

<table>
<thead>
<tr>
<th>Skill</th>
<th>Cronbach’s alfa α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Comprehension</td>
<td>.89</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>.80</td>
</tr>
<tr>
<td>Grammar and Vocabulary</td>
<td>.80</td>
</tr>
<tr>
<td>Writing</td>
<td>.99</td>
</tr>
</tbody>
</table>
Procedural Validity

- The raters were asked to fill in a questionnaire (questions adapted from Cizek & Bunch 2007)
- The raters were very pleased with the organization of the training and they commented that the training day was very useful and necessary
Examinee-centered External Validation

- 13 English teachers assessed the sub-skills of their own students, altogether 290 students.
- The median of the teachers’ assessments correlated quite strongly with the sum score their students got in the matriculation exam: 0.78.
- The teachers’ assessments of the level of their students corresponded also quite well with the levels defined by the test-centered method used by the panel, especially at the levels B1/B2.
4. Results

Research Questions

1. What is the level of the items in the English test?
Table 3. The CEFR levels of the items in the English test (spring 2012)

<table>
<thead>
<tr>
<th>Skill</th>
<th>Items</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>Above B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening 30 items</td>
<td>%</td>
<td>13,3</td>
<td>50,0</td>
<td>36,7</td>
<td></td>
</tr>
<tr>
<td>Reading 30 items</td>
<td>%</td>
<td>20,0</td>
<td>43,4</td>
<td>36,7</td>
<td></td>
</tr>
<tr>
<td>Vocabulary and Grammar 40 items</td>
<td>%</td>
<td>15,0</td>
<td>50,0</td>
<td>15,0</td>
<td>20,0</td>
</tr>
<tr>
<td>Items/100</td>
<td>%</td>
<td>16,0</td>
<td>48,0</td>
<td>28,0</td>
<td>8,0</td>
</tr>
</tbody>
</table>
Table 4. The CEFR levels and the Finnish curriculum levels of the items in the English test

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>30 items</td>
<td>%</td>
<td>13,3</td>
<td>50,0</td>
<td>36,7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>13,3</td>
<td>23,3</td>
<td>27,7</td>
<td>23,3</td>
<td>13,3</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>30 items</td>
<td>%</td>
<td>20,0</td>
<td>43,4</td>
<td>36,7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>6,7</td>
<td>13,3</td>
<td>26,7</td>
<td>16,7</td>
<td>30,0</td>
<td>6,7</td>
</tr>
<tr>
<td>Vocabulary and</td>
<td>40 items</td>
<td>%</td>
<td>15,0</td>
<td>50,0</td>
<td>15,0</td>
<td></td>
<td></td>
<td>20,0</td>
</tr>
<tr>
<td>Grammar</td>
<td></td>
<td>%</td>
<td>15,0</td>
<td>30,0</td>
<td>20,0</td>
<td>10,0</td>
<td>5,0</td>
<td>20,0</td>
</tr>
<tr>
<td>Items/100</td>
<td>%</td>
<td>16,0</td>
<td>48,0</td>
<td>28,0</td>
<td></td>
<td></td>
<td>8,0</td>
<td></td>
</tr>
<tr>
<td>% items</td>
<td>%</td>
<td>2,0</td>
<td>14,0</td>
<td>27,0</td>
<td>21,0</td>
<td>20,0</td>
<td>8,0</td>
<td>8,0</td>
</tr>
</tbody>
</table>
Research Questions

2. Which levels do the students achieve in the English test?
Figure 2. Achieved levels: % of the students

- A2: 0.9%
- B1: 29.6%
- B2: 68.2%
- Above B2: 1.4%
Figure 3. Achieved levels: % of the students

- A2.1: 0.4%
- A2.2: 0.5%
- B1.1: 5.2%
- B1.2: 24.4%
- B2.1: 57.5%
- B2.2: 10.7%
- Above B2.2: 1.4%
3. What is the correspondence between the grades given in the Finnish matriculation exam and the CEFR levels?
**Table 4. Comparison between the grades and the levels**

<table>
<thead>
<tr>
<th>Grade in the Matriculation Exam</th>
<th>Approbatur</th>
<th>Lubenter approbatur</th>
<th>Cum laude approbatur</th>
<th>Magna cum laude approbatur</th>
<th>Eximia cum laude approbatur</th>
<th>Laudatur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of the Finnish version</td>
<td>Low B1.2</td>
<td>Strong B1.2</td>
<td>Low B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.2</td>
</tr>
<tr>
<td></td>
<td>Low B1.2</td>
<td>Strong B1.2</td>
<td>Low B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.2</td>
</tr>
<tr>
<td></td>
<td>Low B1.2</td>
<td>Strong B1.2</td>
<td>Low B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.1</td>
<td>Strong B2.2</td>
</tr>
</tbody>
</table>


5. Conclusion
Further development of the Finnish Matriculation Examination Tests

Development of

1. the writing of the test
2. the grading: report both CEFR levels and indigenous grades
3. a computer-aided test format
4. a new oral subtest as part of the computer-aided test format
Further development of CEFR scales

Drawing on work with the CEFR in Finland at lower secondary school (Tuokko 2009), upper secondary school (Juurakko-Paavola & Takala 2013) and tertiary level (Juurakko & Takala 2001; Juurakko-Paavola 2011) I suggest that

1. The CEFR levels are too broad for capturing well enough the progression of language proficiency,

2. It may be difficult to define verbally descriptors that would reflect fine-grained progression (CEFR 2001, 31-33)

3. Statistical analyses might provide a way to indicate more fine-grained progression
Further development of CEFR scales

4. Three empirically determined sub-levels might be set up:

- low level X (-40 % correct)
- intermediate/typical level X (41-75 % correct)
- strong level X (above 75 % correct)
References


