Replicating results from a CEFR test comparison project across continents

Jamie Dunlea
Society for Testing English Proficiency

Neus Figueras
Generalitat de Catalunya
A pilot study to explore the feasibility of replicating standard setting carried out for an EFL testing program in Japan in a European context
I. Provide further evidence to validate the results of the original standard setting

II. Further our understanding of how far the familiarization and training activities in the Manual are able to build an understanding of the CEFR for teachers working in an EFL context outside of Europe

III. Demonstrate in practice a form of collaboration across contexts which may be able to increase the generalizability of results from individual standard setting projects in relation to the CEFR
The Manual: Building an Argument

- Familiarization
- Specification
- Standardisation training / benchmarking
- Standardisation
- Validation

**CURRENT RESEARCH**
Designed to provide more information about the appropriacy of main standard setting projects in Japan
Background: EIKEN-CEFR Project

- Grade 1: Advanced
- Grade Pre-1
- Grade 2
- Grade Pre-2
- Grade 3
- Grade 4
- Grade 5: Beginner

International admissions to graduate and undergraduate programs

MEXT benchmarks for high school graduates

Grade 3 is a MEXT benchmark for junior high school graduates
### Background: EIKEN-CEFR Project

<table>
<thead>
<tr>
<th>CEFR</th>
<th>EIKEN GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>1</td>
</tr>
<tr>
<td>B2</td>
<td>Pre-1</td>
</tr>
<tr>
<td>B1</td>
<td>2</td>
</tr>
<tr>
<td>A2</td>
<td>Pre-2</td>
</tr>
<tr>
<td>A1</td>
<td>3, 4, 5</td>
</tr>
</tbody>
</table>

Based on specifications stage and analysis of real world test usage (*EIKEN Website*)

Empirical evidence from standard setting (*Dunlea & Matsudaira, 2008; Dunlea & Matsudaira, 2009*)

Data from 2010 standard setting workshops not yet analyzed
Kane (1994, 2001) suggests replicating standard setting with different methods as one source of external validity evidence.

Using different methods and participants, “would provide an especially demanding empirical check on the appropriateness of the cutscore” (Kane, 1994)

If “the Angoff method were used in the original study, the new study might involve an examinee-centered method” (Kane, 1994)
Berk (1986) lists practicability among important criteria for evaluating standard setting methods, and Cizek & Bunch (2007) also emphasise taking note of the “amount of resources available.”

The present study is a pilot study being conducted on different continents, and is completely dependent on the goodwill of teachers and learners who agree to participate.

THIS MEANS IT IS IMPORTANT TO PRIORITIZE PRACTICALITY AND EFFICIENCY, AND REDUCE THE BURDEN ON PARTICIPANTS AS MUCH AS POSSIBLE WHILE OBTAINING A USEFUL AMOUNT OF DATA.
Pilot replication design

- Examinee-centered method: Contrasting groups
- Features of method (from Cizek & Bunch, 2007)
  - Judges classify each examinee as a master/nonmaster in relation to the characteristic which will be tested.
  - Examinees take the test, and two score distributions are formed, one for masters, one for those classified as nonmasters
  - A cutscore is derived that "distinguishes between the two groups"
Pilot replication design

- Focus on one level only, EIKEN Pre-1
- Focus on the Vocabulary and Reading sections of the test only
- do not require equipment for listening or raters for marking writing samples
- allow administration in normal classroom time
Participants: teachers

- Jaeger (1991) concluded that “expert judges should be well experienced in the domains of expertise we demand of them.”
- For this study, we suggest teachers (judges):
  - Should be experienced EFL teachers
  - Should have knowledge of the content and purpose of the CEFR and the Common Reference Levels
  - Should be familiar with the learners they will be rating, and in particular the students’ reading ability
Learners in B2-level classes in language schools in Catalonia, Spain, preparing to take a separate B2-level test

Not all learners will have reached a “mastery” level, that is B2-level of ability in reading, allowing classification into two groups: of masters (B2 level) and nonmasters (below B2 level)

Participants: students

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>Number of teachers</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6</td>
<td>170</td>
</tr>
</tbody>
</table>
The judgment task

- This study also adopts this approach, but with some modifications:
  - In this study, judges rate students they are familiar with.
  - In this study, judges make a holistic mastery/nonmastery decision for each candidate, rather than a separate rating against each descriptor.
The judgment task

- Judges provided with a list of B2-level CEFR descriptors for reading:
  - Overall Reading Comprehension: 1
  - Reading Correspondence: 1
  - Reading for Orientation: 2
  - Reading for Information & Argument: 3
  - Reading Instructions: 1
The judgment task

- Do you consider this learner to be at a sufficient level of proficiency to be able to do the tasks described in the list of CEFR descriptors most of the time in normal circumstances?

- Three alternatives:
  - B2
  - Below B2
  - Cannot judge
Problem: interpreting results

- It is now widely recognized that different methods will result in different cutoffs (Cizek, 2001; Hambleton, 1978; Kaftandjieva, 2010; Zieky, 2001).

- For this reason, Cizek & Bunch (2007) recommend against using multiple methods due to the problems of dealing with the different results that will almost inevitably be obtained.
### Problem: interpreting results

<table>
<thead>
<tr>
<th>EIKEN GRADE</th>
<th>Point of Interest</th>
<th>Pass/fail point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1</td>
<td>B2</td>
<td>70%</td>
</tr>
</tbody>
</table>

- Cutoffs by section from original standard setting (unweighted raw scores)

<table>
<thead>
<tr>
<th></th>
<th>Vocab</th>
<th>Reading</th>
<th>Total</th>
<th>Cutoff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1/B2</td>
<td>14.5</td>
<td>9.5</td>
<td>24</td>
<td>59%</td>
</tr>
<tr>
<td><strong>Total items</strong></td>
<td>25</td>
<td>16</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Overview of Descriptive Statistics for Sample in Current Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Items</strong></td>
</tr>
<tr>
<td><strong>Test takers</strong></td>
</tr>
<tr>
<td><strong>Mean (raw)</strong></td>
</tr>
<tr>
<td><strong>SD</strong></td>
</tr>
<tr>
<td><strong>Min</strong></td>
</tr>
<tr>
<td><strong>Max</strong></td>
</tr>
<tr>
<td><strong>Reliability (α)</strong></td>
</tr>
<tr>
<td><strong>SEM</strong></td>
</tr>
<tr>
<td><strong>Mean item facility</strong></td>
</tr>
<tr>
<td><strong>Mean item discrimination</strong></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>B2 (Master)</th>
<th>Below B2 (Nonmaster)</th>
<th>Cannot Judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>Test takers</td>
<td>101</td>
<td>59</td>
<td>10</td>
</tr>
<tr>
<td>Mean (raw)</td>
<td>32.83</td>
<td>27.25</td>
<td>30.40</td>
</tr>
<tr>
<td>SD</td>
<td>3.86</td>
<td>4.93</td>
<td>3.81</td>
</tr>
<tr>
<td>Min</td>
<td>15</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Max</td>
<td>40</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Median</td>
<td>33</td>
<td>27</td>
<td>30</td>
</tr>
</tbody>
</table>
Cizek & Bunch (2007) suggest several ways for determining cutoff between masters and nonmasters:

- Plot point of overlap between score distributions
- Midpoint between medians of distributions
- Midpoint between means
- Logistic regression (e.g. raw score point at which examinees classified as nonmasters first reach a 50% chance of being classified as masters)
Results

Distribution of raw scores

- Frequency
- Raw Score
- B2
- Below B2
Results

Smoothed B2 / Below B2 Distributions

- B2
- Below wB2
## Results

<table>
<thead>
<tr>
<th>STEP</th>
<th>-2 Log Likelihood</th>
<th>Cox &amp; Snell R-Square</th>
<th>Nagelkerke R-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>158.478</td>
<td>.278</td>
<td>.380</td>
</tr>
</tbody>
</table>

- **Model Chi Square (1)=52.174, *p<.01,**
- **.50=-8.802+.307(x)**
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Cutoff (score)</th>
<th>Cutoff (%)</th>
<th>Pass mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection of distributions</td>
<td>28</td>
<td>68%</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>30.04</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>30</td>
<td>73%</td>
<td>70%</td>
</tr>
<tr>
<td>Regression</td>
<td>30.30</td>
<td>74%</td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>Cutoff=28</th>
<th></th>
<th>Cutoff=30</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Below cutoff</td>
<td>1</td>
<td>31</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Cutoff or higher</td>
<td>9</td>
<td>28</td>
<td>98</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>59</td>
<td>101</td>
<td>10</td>
</tr>
<tr>
<td>Misclassification</td>
<td>19%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

I. Provide further evidence to validate the results of the original standard setting

- The data provide some support for the main claim that Pre-1 certificate holders can be considered to have displayed a B2-level of performance.

- However, the results from this study produce cutoffs much closer to the pass/fail line for certification, raising the possibility of false positive classifications in relation to the CEFR level for borderline test takers.

- At the same time, the higher number of false negatives for the higher cutoff point (30) compared to the lower cutoff point (28) may indicate that a lower cutoff would be more appropriate. As Cizek & Bunch (2007) note, this is a policy decision which needs to take account of the relative impact of the various alternatives (more weight on reducing false negatives or false positives).
II. Did the familiarization and training activities in the Manual build an understanding of the CEFR consistent with that held by colleagues working in a European context

- The higher cutoffs resulting from the present study could suggest that the two groups of teachers (Japan and Spain) have slightly different interpretations of what the CEFR B2 level reading descriptors mean in practice (as test items)

- However, it is also possible that this is an artifact of the different judgment tasks used in the different methods, or indeed of this particular group of learners and teachers. More data and research is needed to make any further inferences.

- There could be an effect for DIF for the test itself, with certain types of items being easier / harder for learners in Japan.
III. Demonstrate in practice a form of collaboration across contexts

- In the context of validating primary standard setting studies, it may not be necessary to replicate all aspects of the original process. Achieving a balance between a practically realizable design and sufficient amounts of data to make useful inferences is important.

- Replicating primary standard setting studies through collaboration with colleagues in other contexts has the potential to greatly add to our knowledge of the consistency of interpretations of the CEFR across contexts and the generalizability of results from individual studies.
“To summarize there is no gold standard, there is no true cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. In three words, nothing is perfect.”
(Kaftandjieva, 2004)
THANKS!

This study was made possible only through the generous offer of collaboration and cooperation by Dr Neus Figueras.

And especially to the teachers and students who volunteered their time and effort...

Thank you very much!


