Collaboration and symbiosis: Using statistics and expert judgement in rater training/standardisation materials

Evelina D. Galaczi
University of Cambridge ESOL Examinations

EALTA 2010
Background

- Cambridge ESOL: Face-to-face speaking assessment
  - A complex process
  - Human raters (oral examiners)
  - 15,000 trained oral examiners worldwide
- Human ratings subject to various sources of error
  - Rater leniency/harshness, central tendency, inconsistency, halo effect
- Rigorous rater training/standardisation is crucial
The role of rater training

- Minimise rater effects
- BUT recognising the fact that rater training cannot make raters into duplicates of each other
- HOW:
  - Familiarisation with assessment scales
  - Benchmarking - examples of speaking test performances and commentaries justifying each mark
  - Applying the assessment scales
An example:

Test 1 (full test): Rino and Gabriela

<table>
<thead>
<tr>
<th>Candidates' Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rino</td>
</tr>
<tr>
<td>Grammar and Vocabulary</td>
</tr>
<tr>
<td>Discourse Management</td>
</tr>
<tr>
<td>Pronunciation</td>
</tr>
<tr>
<td>Interactive Communication</td>
</tr>
<tr>
<td>Gabriela</td>
</tr>
<tr>
<td>Grammar and Vocabulary</td>
</tr>
<tr>
<td>Discourse Management</td>
</tr>
<tr>
<td>Pronunciation</td>
</tr>
<tr>
<td>Interactive Communication</td>
</tr>
</tbody>
</table>

Why were the candidates awarded these marks?

After watching the video use the detailed performance commentaries on the following pages to answer any questions you have about why particular parts were awarded these marks.
So ...

- How do we arrive at standardisation marks?
Stage 1: Marking

- Recorded video performances are marked by oral examiners
- Marking of full tests and test parts
- Rater groups work on combination of exams
Raters:

- **Who?**
  - Past examining experience
  - Good track record from past marking trials (within acceptable parameters for harshness/leniency and internal consistency)
  - Spread geographically

- **How many?**
  - Typically 15 raters marking 20-25 test-takers at each proficiency level

- **Research design?**
  - Fully crossed design
    - E.g., 15 raters x 20 candidates x 4 traits (Grammar and Vocabulary, Discourse Management; Pronunciation; Interactive Communication) at each level
Stage 2: Analysis
Determining candidate marks

- Possibilities
  - One mark ➔ Expert judgement only
  - Multiple marks ➔ Average score
  - Multiple marks ➔ Multi-Facet Rasch Measurement

- BUT…
  - Judgement is subjective
  - Average score does not take into account variation in harshness/leniency of raters, tasks, assessment criteria
Why Multi-Facet Rasch Measurement?

- Provides answers to **individual-level rater effects**
  - Which raters are rating more severely than others?
  - Which raters are inconsistent in using the rating scales?
  - Etc…

- Provides **Fair Average marks** for each candidate
  - E.g., 3.237 ➔ rounded up or down to match marking scheme
Stage 3: Review of rater performance

- Detecting rater effects, e.g.,
  - Harshness/Leniency
  - Consistency: infit/outfit mean square
Misfitting raters

- Misfitting raters removed
- When to stop?
  - “Removing misfit is like cleaning a window. After you clean it the first time, you can see the faint smudges you missed the first time. Then you clean them. Now you can see the fine scratches in the glass. So you polish them out. Now you can see the pane of glass is not exactly flat. So you remedy that. Now you can see that the glass is not all equally transparent. So you ... Following this process, you will finally have no glass left at all!” (Linacre, 2007, personal communication)
  - Raters with mean squares greater than mean + (2xS.D.) are removed from the dataset and analysis is re-run
Stage 4: Review of candidate performance

- Review of:
  - Fair Average mark
  - Mode
  - Distribution of marks

- General rule - Accept the FACETS Fair Average mark as a starting point
Stage 5: Justification of marks

- Group of experienced oral examiners write commentaries to justify each Fair Average mark
  - commentaries include references to the assessment scales and examples from the candidate output
- Commentary writers can query any Fair Average marks which are difficult to support
  - Justification for why a mark should be raised/lowered should be given; these are then reviewed by internal team
What’s a Grammar and Vocabulary Band 3?

- Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms.
- Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics.
Does this describe all Band 3s?

“The search for the perfect scale (the ‘holy scale’) is futile, since … [it] is an artefact that exists to constrain the complexity of the task faced by the raters, to make generalisations, and hence to allow reporting in a way that permits comparison between performances.” (Lumley, 2005, p. 301)
Step 6: Review of marks (again)

- Queried marks (by commentary writers) are reviewed by internal team
  - Data consulted: Actual (i.e., unrounded) Fair Average; Mode; Mean; Frequency distribution of marks
  - To change or not to change?
    - Depends on statistical support
    - 80% of writer queries were supported by the stats ➔ the value of expert judgement
    - For example …
CPE (C2), Lisette, Grammatical Resource, Band 1.5

“Control is better than both Alex and Begum who also have 1.5. Can this be raised?”

Fair average 1.74 ➔ rounded to Band 1.5

**Frequencies**

<table>
<thead>
<tr>
<th>Band</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1</td>
<td>2</td>
</tr>
<tr>
<td>Band 1.5</td>
<td>1</td>
</tr>
<tr>
<td>Band 2</td>
<td>6</td>
</tr>
</tbody>
</table>
BEC H (C1), Mirislava, Lexical Resource, Band 3
“Too many basic errors for this mark. Can it be lowered?”

Fair average 3.08 ➔ Rounded to Band 3

*Frequencies*

<table>
<thead>
<tr>
<th>Band</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 2.5</td>
<td>1</td>
</tr>
<tr>
<td>Band 3</td>
<td>4</td>
</tr>
<tr>
<td>Band 3.5</td>
<td>2</td>
</tr>
<tr>
<td>Band 4</td>
<td>1</td>
</tr>
</tbody>
</table>
Stage 7 Commentary review

- The commentaries are edited with two main questions in mind:
  - Do they reflect the performance descriptors in the assessment scale for the mark given?
  - Is there some consistency across the tests/exams in terms of commentaries for the same marks?

- **Editing team** check commentaries and performances arranged in score order for each criterion

- **Speaking Test ‘Chair’** checks commentaries and performances in the order in which they will be watched by oral examiners during standardisation
Finally … Lessons learned

- The importance of quality checks
  - Thousands of OEs with different perspectives
  - Need rigorously chosen and justified standardisation marks
- Familiarity with Multi-Facet Rasch Measurement and FACETS to allow for operational use
- Initial resistance in “selling” idea of statistical evidence to oral examiners
The moral of the story

SYMBIOSIS

STATISTICAL EVIDENCE

EXPERT JUDGEMENT
Thank you!

Evelina

Galaczi.e@cambridgeESOL.org