Looking beyond scores – a study of raters and ratings of speaking

Linda Borger
University of Gothenburg

The 11th Ealta Conference
30 May, 2014
University of Warwick
Overview

1. Background
2. Research questions
3. Method of data collection and data analysis
4. First results and observations
Background

National tests in Sweden
• Enhance equity and fairness; compulsory
• Marked by teachers
• Reliability questioned by Sw Schools Inspectorate – however, oral parts not examined

Relation between Swedish foreign language syllabuses and the CEFR
• Based on a communicative approach
• Entrance/Pass levels intended to correspond to the CEFR
• Extensive textual analyses; no empirical validation yet
Aim of study

Rating of a paired speaking test
– decision-making
– scores

Secondary aim
– a small-scale, tentative comparison between Swedish performance standards and CEFR-levels
Research questions

RQ 1: Decision-making processes
• What features of a candidate’s performance are salient to raters?

RQ 2: Rating/scores
• What can be noticed regarding possible rater profiles and variability?

RQ 3: Relation between scores and rater comments
• What is the possible relationship between scores and rater comments?

RQ 4: Comparison CEFR
• What levels in the CEFR do external raters judge the performances of the Swedish students to be?
Data collection

Participants

- Swedish raters
  teachers of English at upper secondary level (n = 17)
- External CEFR raters
  Finland (n = 7) and Spain (n = 7)

Audio-rating of six conversations (12 performances)
- Authentic material from Sw national test of English at upper secondary level
- First part: focus on oral production (short text to read and summarize); some interaction
- Second part: focus on oral interaction (discussion and argumentation)
- A boy and a girl in each conversation
- Examiner has a minimal role
Data collection and analyses

One-day seminar

- Rater notes (while listening)
- Summary comments (after listening)
- Scores

‘Qualitative’ data

Content analysis: segmentation and coding using NVivo 10 (qualitative data analysis computer software)

‘Quantitative’ data

- Descriptive statistics
- Correlations
- Internal consistency
Two different scales

Sweden: Ten-point scale tentatively adapted to the CEFR. Entrance level intended to correspond to a minimal B2

<table>
<thead>
<tr>
<th></th>
<th>F-</th>
<th>F+</th>
<th>E-</th>
<th>E+</th>
<th>D-</th>
<th>D+</th>
<th>C-</th>
<th>C+</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Finland/Spain: Nine-point scale in accordance with the CEFR (i.e. the full range) from the Manual (2009)

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A2+</th>
<th>B1</th>
<th>B1+</th>
<th>B2</th>
<th>B2+</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
RQ 2: What can be noticed regarding (rater profiles and) variability?

Swedish raters: median och range

N = 17
Results

Statistics for Swedish raters

RQ 2: What can be noticed regarding (rater profiles and) variability?

Mean values for Sw raters

N = 17
Results and observations: Swedish raters

Correlations between Swedish raters’ scores

• Spearman’s rank-order correlation: 0.59 – 0.95
• Kendall’s Tau correlation: 0.47 – 0.89
• Cronbach’s alpha: 0.98

Some observations about Swedish raters

• Reasonably good rater agreement and consistency
• Rater profiles - differences in degrees of leniency and severity (mean of raters varies between 5,58 – 8,0)
• Some performances obviously more difficult to agree on
Results

Statistics for European raters

RQ 4: Comparison CEFR

Mean values for European raters

N = 14
Observations: European raters

RQ 4: What levels in the CEFR do external raters judge the performances of the Swedish students to be?

• Almost identical rank ordering among European and Swedish raters
• Similar variability as in the Swedish group
• The mean values for the European raters are between B1+ and B2+ (in line with intention of test)
• European raters somewhat harsher around the cut-off point (cf. minimal level)
CEFR: Relevant Qualitative factors for production and spoken interaction (adapted from the Manual, Table A4 and A5, p. 148-149)

- **LINGUISTIC RANGE**
  - General linguistic range (incl. express viewpoints)
  - Vocabulary range

- **LINGUISTIC ACCURACY**
  - Grammatical accuracy
  - Vocabulary control
  - Phonological control

- **SOCIO-LINGUISTIC**
  - Socio-linguistic appropriateness

- **PRAGMATIC - FLUENCY**
  - Fluency, Flexibility

- **PRAGMATIC - COHERENCE**
  - Coherence and cohesion
  - Thematic development (Topic development)

- **INTERACTION**
  - Turntaking, Cooperating

- **STRATEGIC**
  - Compensating, Monitoring and Repair
Coding scheme: rater comments

RQ1: What features of a candidate’s performance are salient to raters?

Criterion features
- Accuracy
- Coherence
- Fluency
- Interaction
- Range
- Strategies
- Socio-linguistic
- Summary of text
- Intelligibility
- Length of response
- Overall

Non-criterion features
- Evaluative response
  - Positive
  - Negative
  - Mixed
- Focus of response
  - Inter/Intra-candidate comparison
- Rater reflection
Example: Coding of rater comment
Low-scoring student (Sw rater)

Grade: F+
Sums up his card reasonably well Summary of text/Pos
and gives his thoughts. Range/Pos
He invites his partner to respond throughout Interaction/Pos
and gives examples to support his point. Coherence/Pos
However, due to poor linguistic resources/pronunciation Range +
Accuracy/Neg
it lacks clarity in a number of places. Coherence/Neg
Only towards the end of the test does he improve and convey the
message without any obstacles. Coherence/Pos, Inter-candidate
comparison over time
/.../
Due to the fact the message was impeded in a number of places I feel I
cannot pass this pupil even though he was communicative. Rater
reflection
Salient features – Swedish raters

1. Accuracy
2. Range
3. Coherence
4. Fluency
5. Interaction

Coded categories for 10 Sw raters

- Range: 176
- Accuracy: 185
- Fluency: 108
- Interaction: 100
- Coherence: 141
- Strategies: 41
- SLAPP: 7
- Summary: 56
- Intelligibility: 12
- Length of Response: 0
- Overall: 16
Salient features – European raters

1. Range
2. Accuracy, fluency
3. Interaction
4. Coherence

Coded categories for 10 European raters
Evaluative response Swedish raters
Evaluative response European raters

Evaluative responses 10 European raters

Mixed  Negative  Positive
### Relation between comments and scores

**RQ 2: What is the possible relationship between scores and rater comments?**

*(Candidate 6 male student)*

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>C-</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACCURACY</strong></td>
<td>Makes quite a few subject verb agreement errors.</td>
<td>&quot;Linguistic insecurity at times&quot;</td>
<td>Grammar, vocabulary, pronunciation were all very good, natural feel for English</td>
</tr>
<tr>
<td></td>
<td>Somewhat unidiomatic in places, e.g. elder people, relaxation for relaxation, the ways of relax, it’s more individually.</td>
<td>&quot;Some grammatical errors but they do not cause misunderstanding&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Good pronunciation on the whole.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RANGE</strong></td>
<td>Displays a relatively good range of language (e.g. research, motivation to work, skill), but greater range required.</td>
<td>&quot;Some really good phrases and expressions&quot;</td>
<td>could present, discuss and reflect to some degree on his answers.</td>
</tr>
<tr>
<td><strong>INTERACTION</strong></td>
<td>Is interactive throughout the test.</td>
<td>&quot;Asks questions and moves the conversation forward&quot;</td>
<td>He had very good communication skills</td>
</tr>
</tbody>
</table>
Thank you for your attention!

E-mail: linda.borger@gu.se
Selected References


