Paper presented at CALA 2017 Symposium

Speaking proficiency is in the eye of the beholder? A comparison of justifications from university instructors and experienced raters

Alex Volkov, Zhi Li, & Jayanti Banerjee

Paragon Testing Enterprises

Agenda

- ✓ Background
- ✓ Research Questions
- ✓ Methodology
- ✓ Findings
- ✓ Discussion

Background

Rating Scale Development

- ✓ Multistage project, involving qualitative and quantitative stages
- ✓ Methodology similar to North and Schneider (1998)
- ✓ First, wide range of descriptors
- ✓ Later, narrowing down and organizing the list of descriptors

Background

Rating Scale Development

- ✓ Key question is construct representation
- ✓ First, including various sources of data
- ✓ Second, incorporating different perspectives – linguists, raters, decision-making institutions
- ✓ Different perspectives are often combined, but rarely investigated

Some studies on raters' and university professors' judgment

Professional raters

- ✓ Writing: Barkaoui (2010)
- ✓ Speaking: Winke & Gass (2013); Winke et al. (2013); Wei & Llosa (2015)

University instructors

- ✓ Writing: Huang & Foote (2010); Vann et al. (1984); Song & Caruso (1996); Janopoulos (1992); Zhu (2004)
- ✓ Speaking: Elliot & Hickam (1997); Zhang & Elder (2011)

TESTING ENTERPRISES Paragon

Research Questions

1. To what extent did the two groups of panelists judge the speech samples differently?

2. To what extent did the two groups of panelists differ in justifying their ratings?

Data

- ✓ Academic language test under investigation
- ✓ Pilot structure
- ✓ Each test-taker produced independent tasks, speaking tasks based on listening and based on reading

Methodology

- ✓ 50 responses
- ✓ Five experienced raters and three experienced professors
- ✓ Overlapping pattern – 20 responses per panelist
- ✓ Rated *Not ready/Ready/Above ready* for Canadian post-secondary level
- ✓ Provided justifications and explanations

Methodology

- ✓ Many-Facets Rasch Measurement
- ✓ Focus is on construct definition, understanding target language performance
- ✓ Therefore, qualitative justifications are the key source of data
- ✓ Singled out all the descriptors and statements. Categorized the descriptors

Findings (shared)

✓ **Errors are viewed through overall comprehensibility and message**

- *When the mistakes are made they do not seriously impede comprehension* (Rater 2)
- *Problems with grammar and vocabulary that consistently impede comprehension* (Professor 2)

✓ **Self-correction is seen as an important positive feature**

- *Does some self-correcting to clarify/correct his message which indicates that he is monitoring his message* (Rater 4)
- *Does not demonstrate awareness of these mistakes, since she does not pause for grammatical and lexical repair* (Professor 3)

✓ **Understanding the task and completing the task are seen as two different categories**

- *Demonstrates understanding of the question itself* (Professor 1)
- *Indicates good understanding of the prompt* (Rater 4)

Findings (raters)

✓ **Always splitting language into analytical categories –lexical and syntactic features, pronunciation, precision, etc.**

✓ **Complexity and sophistication of language**

- *There are points of sophisticated grammar usage* (Rater 1)
- *The response is simple, it lacks complexity* (Rater 4)

✓ **Depth and complexity of ideas expressed**

- *Ideas convey a deeper meaning* (Rater 5)
- *Ideas are general and elementary in thought* (Rater 3)

✓ **More levels of performance (potential for diagnostic information)**

- *Her communication is clear, but weak in quality* (Rater 3)
- *I would suggest that further practice in oral language would be beneficial* (Rater 4)

Findings (professors)

✓ **Language is often discussed more holistically**

- *I don't think her level of English would allow her to….* (Professor 1)
- *Did not demonstrate much breadth of English skills* (Professor 2)

✓ **Answering the question is often seen as the primary category**

- *I rated this student as 'not ready' because she does not answer the question correctly in any way* (Professor 1)

✓ **Relevance and accuracy of the information**

- *Focused only on one point of the lecture, which wasn't the main one* (Professor 3)
- *He brought in points that were not relevant* (Professor 1)
- *The student missed that fact that there are two teaching assistants* (Professor 1)

✓ **Rephrasing the information in own words**

- *Phrases seem to be directly taken from the passage* (Professor 2)
- *Makes a solid attempt to answer the question in her own words* (Professor 1)

Findings (overall)

- ✓ Considerable overlap in principles and criteria
- ✓ The groups are not homogeneous – we have noticed individual differences on views and approaches
- ✓ Raters: holistic meaning AND analytic language. Professors: task, then language
- ✓ Professors: often a binary decisions. Raters: often a scale/gradation

Questions and Dilemmas

- ✓ Grid-like scales (raters analytical thinking) vs EBBS (Turner and Upshur, 1996) – professor's step-by-step thinking
- ✓ Accuracy of source information in integrated skills. Is this indeed part of the construct?
- ✓ Who is the ultimate judge of the construct?

Thank you!

Speaking proficiency is in the eye of the beholder?
A comparison of justifications from university instructors and experienced raters

Questions?