Title of Project:
Rater Expertise in a Second Language Speaking Assessment: The Influence of Training and Experience

Researcher:
Larry Davis
University of Hawai’i, Mānoa
larrydavis98@yahoo.com

Research Supervisor:
Dr. John Norris

Project Summary:

In tests of speaking it is the rater who translates a performance into a score; it is therefore the rater who is the final arbiter of what a given score means. Accordingly, to understand the meaning of a score we must understand how and why raters make particular scoring decisions. While considerable effort has been devoted to the analysis of raters’ scoring patterns, many questions remain regarding how people learn to make judgments of complex phenomena such as language ability. For example, it has been reported that as essay raters gain experience they tend to be more internally consistent but do not necessarily become more consistent with other raters (Lim, 2009; Weigle, 1994, 1998). However, it is less clear how decision making changes with training or experience to actually produce such increased intra-rater reliability. Understanding how raters develop the expertise to reliably and accurately score speaking responses has obvious implications for the design of rater training and scoring procedures.

Moreover, studies in the area of rater decision making tend to be descriptive in nature, and within speaking assessment little has been done to investigate the basic mechanisms behind raters’ decision making. A common assumption seems to be that scoring judgments are made by applying a scoring rubric to a speaking performance, which nicely fits the widely cited definition of measurement as the application of a rule to an observation (Stevens, 1946). However, a different view has recently emerged from the field of psychophysics and behavioral economics which asserts that magnitude judgments are made by comparing the observation at hand with other examples in the environment or memory (Laming, 2004; Stewart, Chater, & Brown, 2006). Within this framework judgments are seen as fundamentally relative in nature rather than being the application of an absolute rule.
The International Research Foundation for English Language Education

The overall purpose of this dissertation was to extend our understanding of rater expertise within a speaking test context. The primary focus included an investigation of the ways in which rater scoring patterns and behavior change with training and experience, as well as the characteristics of scoring behavior and cognition associated with raters showing more or less desirable patterns in scoring. In addition, an initial effort was made to examine the possibility that judgments of language ability are relative in nature. The following research questions were examined:

1. What effects do training and experience have on scoring patterns and scoring behavior of raters?
2. What features of scoring behavior and cognition distinguish more-proficient and less-proficient raters?
3. Can rater decision making be understood in terms of a relative view of judgment?

To examine these questions, 20 experienced teachers of English scored recorded examinee responses from the TOEFL iBT speaking test prior to training and in three sessions following training, over a period of about three weeks. In the first session previously scored example responses were the only materials given to raters; raters awarded scores on the basis of their native scoring criteria along with whatever information they could take from the examples. This was followed by training which included exposure to the scoring rubric, review of example responses with scoring explanations, and practice scoring with feedback given after each score (the feedback being a previously established reference score for the same response). For the final three scoring sessions raters were simply asked to review the scoring rubric and listen to a set of example responses at the start of each session. Raters scored 100 responses in each session along with an additional 20 responses where they verbally reported what they were thinking as they listened to an examinee response and made a scoring decision.

For Research Question 1, scores from each rater were analyzed for consistency in scoring patterns and agreement with other raters, as well as for accuracy (agreement with reference scores). Various aspects of raters' interaction with the scoring materials were also recorded to determine if certain behaviors, such as the time taken to reach a scoring decision, were associated with the reliability and accuracy of scores. Somewhat surprisingly, rater severity and internal consistency (measured via Rasch analysis) were already of a standard typical for operational language performance tests in the first scoring session, before raters had received training or even seen the scoring rubric. This result suggests that the example responses provided in the first session may have been adequate for raters to apply the rating scale, at least for the experienced teachers used as raters in the study. Training played a role as well, however, resulting in increased agreement between raters and improved agreement with established reference scores. Additional experience gained after training appeared to have little effect on rater scoring patterns, although agreement with reference scores continued to increase somewhat.
For Research Question 2, less-, intermediate-, and more-proficient raters were identified on the basis of scoring consistency and accuracy and compared in terms of behavior and thought processes while scoring. More proficient raters reviewed example responses more often and took longer to make scoring decisions, suggesting the possibility that rater behavior while scoring may influence the accuracy and reliability of scores. Analyses of rater’s verbal reports found no obvious differences in the frequency with which raters mentioned various language features while scoring, and considerable individual variation was seen. However, following training, mention of language features related to topic development increased for all groups, probably in response to specific instructions for evaluating topical development provided in the training. The frequency of comments on the scoring process also increased following training, and overall the training appeared to influence both the attention raters gave to certain language features and the scoring process more generally.

For Research Question 3 the extent to which scoring decisions are made by comparing examinee responses was investigated. First, scores produced by each rater were examined for sequence effect, or the tendency for a score to be correlated with the score given to the previous test taker. In the first study, small but significant correlations were found between scores produced in succession, supporting the view that scoring judgments are relative in nature. Second, an experiment was conducted where raters were presented with pairs of responses and asked to decide if the responses deserved the same or different scores; the prediction was that it would be easier to distinguish between similar responses when these could be compared side-by-side. Contrary to this prediction, discrimination between examinees was actually poorer than when the same responses were individually scored. Many raters also felt that the comparison task was cognitively demanding because details of the first response had to be held in mind while judging the second response. This memory load may have been a confounding factor and complicates interpretation of the experiment results, but from a practical standpoint underscores the challenge of judging speaking ability through direct comparison of responses.

The study has a number of implications for assessment practice. First, the results highlight the importance of examples for aligning rater perceptions to the rating scale. With only the exemplars for guidance, raters in the study were able to achieve reasonably good score reliability and accuracy in the first scoring session. So, it may be useful to make examples available while scoring and encourage raters to refer to the examples often. The results also raise the question of the degree to which descriptions of performance found in scoring rubrics actually influence raters’ scores. While exposure to the scoring rubric appeared to influence rater perceptions and improve accuracy in scoring, it seems likely that scoring decisions also rely on the examples used to develop an understanding of the rating scale. Accordingly, scoring rubrics may represent
only a partial description of the scoring criteria actually employed in a language performance test.

The results also demonstrate that training has value, as indicated by the improved scoring performance seen after raters completed a series of typical training activities. Subsequent experience resulted in little further improvement in scoring performance, suggesting that unguided experience may not necessarily be very effective in developing scoring expertise, at least once a rater has reached a certain level of performance. This view is consistent with the broader literature on expertise, which suggests that elite levels of performance can be obtained only through focused practice, and simple repetition is not enough.

Finally, in this study I evaluated a possible mechanism for how scoring decisions are made; namely, that judgments of language ability are relative in nature. While the results of the current study are inconclusive, several findings suggest that further examination of this theoretical position is warranted, including the finding that (a) scoring judgments were influenced by similar examples of performance available in the immediate environment, as indicated by a sequence effect; and (b) greater use of exemplars for comparison was associated with scoring accuracy/consistency in more-proficient raters. Although the empirical findings produced so far are quite modest, the current study introduces to the domain of language assessment a new, and testable, framework for thinking about the way raters make judgments of language ability.
References


[Computer software]. Chicago, IL: Publisher.


University of Melbourne, Melbourne, Australia.


Quarterly, 36, 49–70.


