A Closer Take a Look at Rater Bias in English Teacher Employment Examinations

Tomoyasu Akiyama

Department of English Language & Literature

akitomo@koshigaya.bunkyo.ac.jp

Abstract

This study examines rater variability in employing prospective English teachers using both quantitative (Many-Facet Rasch Analysis, MFRA) and qualitative (think-aloud) methods. A total of 17 raters evaluated 30 candidates based on six assessment criteria, namely, lesson flow, instruction ability, delivery, personality, expertise, and overall employment decision. Data were analyzed using MFRA to identify rater severity and consistency as well as biased interactions between candidates and raters. Think-aloud data were also analyzed to explore raters' use of assessment criteria and the possible reasons for biased outcomes. Results indicate that most of the raters show consistent rating patterns but with various levels of severity; about a quarter of interactions between raters and candidates are significantly biased. The think-aloud analysis shows that raters evaluated candidates not only based on different interpretations of and attention to sub-criteria of the six assessment criteria but also according to the different teaching values of the raters. Finally, this work gives implications for rater training.

Keywords

rater variability, microteaching, bias analysis, Many-Facet Rasch Analysis (MFRA), think-aloud method

1 Introduction

Investigating rater variability is among the most popular research areas in language testing (e.g., Lumley & McNamara, 1995; Eckes, 2008). The development of Many-Facet Rasch Analysis (MFRA) has facilitated the research on rater-mediated assessment, whereas the availability of scores awarded by raters has also benefited score variability studies (Knock, 2011). Rater variability affects test score interpretations as well as the use of test scores on test-takers' consequences. Using both types of data, the present work investigates the raters' use of assessment criteria and factors causing unexpected rater behaviors and biases in an English teacher employment test context.

2 Research Methods

Research questions

The study addresses the following questions to investigate rater variability focusing on interactions between candidates and raters:

RQ 1. To what extent were the raters harsh and consistent among one another?

RQ 2. How did the raters use the assessment criteria?

RQ 3. What were the factors of rater bias in assessing the candidates' teaching performance?

Participants

Twenty university students who were training to become English teachers and 10 in-service teachers participated in this study. All participants were required to introduce one of the following six target grammar points in their microteaching lesson (e.g., *Do you___*? *Can you___*?).

The test procedure included the following steps: The participant was assigned to one of the six target grammar points for which he/she was asked to draft a lesson plan within 20 minutes. After which, the participant carried out his/her lesson plan for about five minutes to demonstrate teaching skills. All demonstrations were videotaped for assessment based on six criteria.

Raters and assessment criteria

17 raters were required to rate 30 candidates individually by viewing videotaped demonstrations. They were asked to dictate their specific observations for each of the six assessment criteria into a digital voice recorder. The recorded statements were analyzed to

investigate rating processes and possible reasons for bias in interactions between candidates and raters. All raters underwent a two-hour training with the researcher; the training covered the purpose of the study and assessment criteria, including a sample run of rating one to two candidates by both the rater and researcher. While administering one training session for all raters was an ideal situation to obtain uniform conditions, it was not achieved owing to raters' schedule constraints.

3 Results

Figure 1 shows the degree of rater consistency (line chart) and rater severity (bar chart). If the infit mean square (IMS) is within the acceptable range (from 0.7 to 1.3, in McNamara, 1996), rating patterns are more or less similar to that expected by the Rasch model. This study employed a narrower range owing to the high-stakes context of teacher employment examinations.

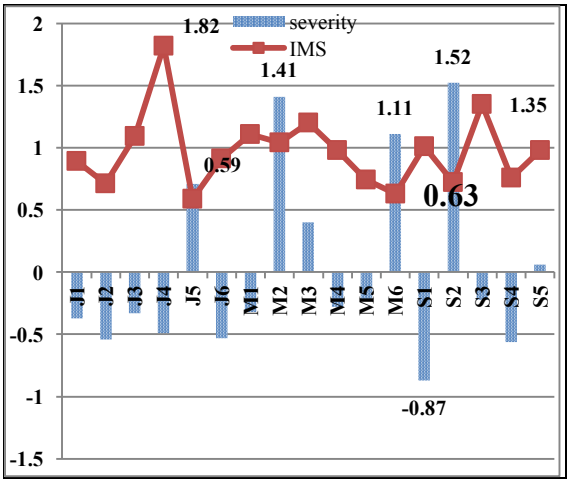

Figure 1 MFRA rater consistency (Infit Mean Square, IMS) and rater severity
Notes: Reliability: 0.98; Separation: 6.8

Figure 2 shows the extent to which the 17 raters used sub-criteria on each criterion and extra-assessment criteria. The left-side bar charts for each rater show the original assessment criteria and the right-side one, extra-assessment criteria. All 17 raters did not use the assessment criteria fully. S5 was the most frequent user of the original assessment criteria (80%) followed by J6, S1, and J2. M6 and S4 preferred the extra-assessment criteria and used only about 30% of the original one. Thus, constructs for assessment based on the results of raters' use of assessment criteria seemed different.

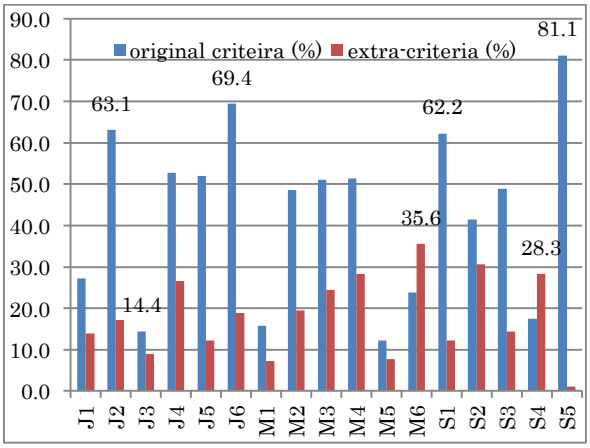

Figure 2 Use of assessment criteria by 17 raters (%)

4 Conclusions

The following is a summary of the present study's main points:

- Most raters evaluated the candidates consistently but with different degrees of severity.
- The raters used assessment criteria differently, applying their own interpretations of the criteria. They also introduced their own assessment criteria apart from those provided.
- Biased interactions might have occurred when a candidate's teaching demonstration was inconsistent with a rater's ideal teaching scenario or core teaching values.

This study investigated rater behavior, employing both quantitative (MFRA) and qualitative (think-aloud method) analyses. The results of the study showed that raters evaluated the teacher candidates with a wide range of assessment criteria use, different interpretations of and attention to sub-criteria, and application of different teaching values.

5 References

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. Language Testing, 25(2) 155–185.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. Language Testing, 28(2) 179–200.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications of training. Language Testing, 12(1) 54–71.

McNamara, T. (1996). Measuring Second Language Performance. London/New York: Longman.