Linking Rater Behavior to Criteria on a Local Paired Speaking Task: A Mixed-Methods Approach

Elnaz Kia

Northern Arizona University

Abstract

This paper reports on a mixed-methods approach to evaluate the rating criteria used for a local paired-speaking task and to examine potential biases between raters and the examinees, L1, and the criteria. I examined 5 raters' ratings on 56 paired speaking speech samples from examinees of Arabic, Chinese, and Portuguese L1. Facets modeled in the analysis were examinee, rater, L1, and criteria. I analyzed the ratings using a multi-faceted Rasch measurement approach (Linacre, 2015). Insights were added to the qualitative results based on the raters' comments during the think-aloud sessions. Results indicated that there was no bias between the raters and the examinees or between the raters and the examinee's L1. However, significant biases were located between the raters and the criteria. Analysis of the scales showed that raters were not using the scale fully and it is suggested to change the 5-points scale into a 3-points scale. Suggestions are made regarding improving the rating criteria and improving rater consistency.

Keywords: FACETS, mixed-methods approach, Item Response Theory (IRT), rater bias, paired-speaking task

Linking Rater Behavior to Criteria on a Local Paired Speaking Task: A Mixed-Methods Approach

Background

With the emergence of communicative approaches to second language learning, there has been growing interest in using paired speaking tasks as a means of assessing interactional competence in classrooms (McNamara, 1996; Taylor & Wigglesworth, 2009). While paired speaking tasks have proved themselves as advantageous over individual tasks or interviews between an examiner and examinee (Nakatsuhara, 2004), there are several challenges in using them concerning language testing. First, due to the complexity of interactional competence, it is difficult to define the constructs on the criteria. Second, alike other performance-based tasks, paired speaking tasks are also subject to rater bias, which threatens the validity of the outcomes. Rater severity or leniency can be investigated in relation to the overall score, as well as a certain construct or criterion (McNamara, 1996).

In contrast to the large number of studies on the interaction between the rater and the criteria used for writing tasks, little attention has been given to oral proficiency tasks, particularly paired speaking tasks. Therefore, in an attempt to fill this gap, the present study aims to measure the validity of a paired speaking rating scale by exploring the link between the rater and the rating scale by applying multi-facet Rasch model, along with qualitative data analysis. Findings of this study may add insights to the selection of constructs appropriate for evaluating a paired speaking task in the intermediate level of proficiency.

Research Questions

1. To what extent do trained raters show consistency and accuracy in awarding scores to examinees using the paired speaking criteria?
2. How well do the rating criteria designed for the paired speaking task represent the targeted construct?
3. Is there a relationship between the examinees' L1 and raters' severity?
4. Is there a relationship between the raters' severity and the level of difficulty of the rating criteria?
5. Is there a relationship between the raters' perceived importance on rating criteria and raters' severity?

Methodology

Participants

The participants were 5 teachers at an IEP (Intensive English Program) in a large university in the United States. Of these 3 were female and 2 were male. Four of the participants were TAs (Teaching Assistants) and one other participant was a lecturer in the IEP. Although two of the participants were non-native speakers of English, they were both fluent, native-like speakers. In terms of teaching experience, all the participants had taught various English skills; however, one of the participants was an expert in writing and had mostly taught writing for all her career. With regards to rating experience, all the participants had prior experience using the rating criteria being investigated in the present study.

Materials

Test taker speech samples. Archived speech samples from the same IEP were used in this study, due to availability of large collection of recorded responses. Overall, sixty six

examinees' speech samples were selected from three levels of proficiency (2, 3, and 4). The students' proficiency in these levels is equivalent to 16-31, 32-44, and 45-56 TOEFL iBT score range respectively. The number of examinees and corresponding speech files that were scored in each level were as follows: 16 examinees (8 speech files) from level 2, 19 examinees (10 speech files) from level 3, and 21 examinees (11 speech files) from level 4. Regarding the examinees' L1s, there were 40 Arabic speakers, 8 Portuguese speakers, and 8 Chinese speakers.

Prompts. In order to control for prompt difficulty, the speech samples were selected from paired speaking tasks with similar prompts across levels. Despite the minor differences in the prompts, it is believed that the slight difficulty or easiness of each prompt was in concordance with ability of the examinees.

Rating criteria. The rating criteria investigated in this study was an analytic rubric which was currently being used for level 3 students at the IEP and comprised of three sub-rating criteria: collaboration, task completion, and style. On each criterion, examinee performance was scored on a scale of 0 to 5 (0, 1, 3.25, 3.75, 4.25, and 5).

Rating criteria instruction. A handout was prepared by the researcher, including overall explanations about the rating criteria, as well as detailed information about each sub-rating category and its descriptors. Multiple examples were provided for some of the descriptors to make them more concrete.

Think-aloud protocol script. The purpose of the project and the think-aloud procedure were defined in the script. In addition, a review of the training session and the think-aloud procedure were written. Moreover, 5 questions were posed in the script. The first 3 questions which were asked after the training and modeling session were aimed at eliciting raters' general impression of the rubric. The last 2 questions which were asked at the end of the think-aloud

session were aimed at investigating raters' perceived level of severity and also level of difficulty of each sub-rating category on the rating criteria.

Procedure

The training and think-aloud session were combined into one session and occurred subsequently. The rating criteria and its instruction were emailed to the raters a day before their sessions and they were asked to review them briefly. In the training session, after stating the purpose of the study, rater training, and the think-aloud procedure, the researcher went through the instruction with the rater and made sure that the rater understood the instructions well. Afterwards, the rater was asked if he/she had any questions.

After the training step, the raters were asked 3 questions regarding the design, clarity, comprehensiveness of the rubric. Subsequently, 3 speech files assigned for the think-aloud session were played one by one and after finishing each sample the raters started doing the think-aloud. Finally, the researcher asked 2 follow up questions regarding the raters' perception of the level of importance of the sub-rating categories and raters' potential bias towards the categories. All 3 think-alouds, as well as the follow-up questions were recorded by a voice recorder.

The last part of the data collection was individual ratings by the raters, which was performed at their convenience without the presence of the researcher. The rating was partially crossed among the raters. Each rater scored 12 out of 56 examinees, which is about 21% of the whole data. The 12 common speech samples were equally selected from levels 2, 3, and 4. For the rest of speech sample, the researcher assigned samples in a way that there were common ratings between every pair of raters in each level.

Data Analysis

A Many-Facet Rasch Model (MFRM) was used to answer the research questions in this study. To conduct the analysis, I used the FACETS computer program, version 3.71.4 (Linacre, 2015). To match the variables in the study, 4 facets were included in the model: Examinee (n=56); L1 (Arabic, Chinese, and Portuguese); Rater (n=5); Criteria (collaboration, task completion, style). L1 was added as a dummy facet to the analysis in order to explore the potential bias of the raters towards examinees' L1. The rating scale (0, 1, 3.25, 3.75, 4.25, 5) was slightly modified in the model, since not all parts of the scale were being used. Two separate analyses were run using FACETS: 1. fit analysis of all the 4 facets; 2. three bias interaction analyses to determine the potential biased interactions between the raters and the examinees, the raters and examinees' L1, and raters' and the criteria. Selected answers to the research questions were supported by data from the think-aloud procedure and the follow-up question.

Results

Research Question 1

The first question was whether trained raters show accuracy and consistency in awarding scores to examinees using the paired speaking criteria. The severity span between the most severe rater (rater 3) and the least severe rater (rater 4) was 1.96 logits with the mean value of .00 (SD=.74). Mean of .00 suggests that the raters had average level of severity for this group of students. SD of .74 also suggests that raters were clustered .74 logits below and above the average (0.0). While raters 5, 1, and 4 had fairly similar levels of severity, raters 3 and 2 differed from the other raters by higher logit values. Care should be taken reporting these values, since the standard error measure (SE) is high. The significant high reliability (0.93) and strata (5.09)

indices show that the raters were significantly different in severity ($p = 0.0 < .001$). In other words, the raters were not interchangeable (McNamara, 1996).

In order to examine if the raters were internally self-consistent across the examinees, the criteria fit statistics was investigated. Raters 1 and 2 showed a fit value close to the expected value of 1 (1.01 and 0.82 respectively). That is, both raters used the rating scale consistently and maintained their level of severity across different facets. On the other hand, rater 5 (1.60, zstd =2.8) showed significant misfit which signals inconsistency in using the rating scale across examinees and the criteria.

Research Question 2

The second research question was regarding the rating criteria and how well it represents the targeted construct. Results showed that it was the hardest for the students to get high ratings on task completion, with a difficulty of 0.46 logits (SE = .14), and it was the easiest to obtain high ratings on collaboration, with a difficulty of -0.48. The separation index of 3.42 and high reliability of .84 indicate that the analysis reliably distinguishes between more than 3 distinct levels of difficulty among the rating criteria. The fit statistics of the three criteria are within the acceptable range of fit. This suggests that the criteria work well together and they are all part of the same construct.

Research question 3

To investigate whether the raters score particular examinees harsher or more leniently than their usual scoring patterns, the MFRM was run to perform a two-way interaction (i.e., Rater × Examinee) analysis. The bias interaction produced by FACETS included 2 interactions between raters and examinees with potential bias: rater 5 and examinee 14, and rater 4 and examinee 14. Although the t-statistics were beyond absolute 2 (McNamara, 1996), none of the

interactions were significant at the .05 level. For rater 5, the observations for examinee 14 were .85 higher than expected and for rater 4, the observations were .85 lower than expected. In examining the fit statistics, examinee 14 was highly beyond the fit range (3.52, zstd =2.8). Moreover, rater 5 showed misfit (1.60, zstd =2.8). It is recommended that raters receive additional training.

Research Question 4

Research question 4 was whether there is a bias between the raters and the examinee's L1. Only one interaction with potential bias was reported between rater 2 and Portuguese learners. The observation was .36 higher than expected, but the t-statistic was not beyond absolute 2. The chi-square test was also not significant. In other words, there was no bias between the raters and examinee's L1.

Research Question 5

In order to examine the interaction between raters and the criteria, a two-way bias/interaction (i.e., Rater × Criteria) analysis was run to examine whether the raters maintained their rating across three rating criteria (i.e. task completion, collaboration, and style).

The percentage of the Rater × Criteria interactions resulting from differences between observed and expected ratings was low (4 interactions out of 15). All these interactions had absolute t values equal or greater than 2, and they were all significant at the 0.05 level. In other words, raters experienced a level of difficulty applying the criteria in a consistent manner in comparison to their overall level of severity. Two of the interactions were between rater 5 who was a fairly lenient rater and two criteria (i.e. collaboration and task completion). When scoring task completion, she was significantly lenient by 1.44 logit value. However, examining the bias size, t-statistic and the probability for collaboration shows that, rater 5 was significantly severe

by -1.03 logit value. Rater 5's fit values of 1.6 and 1.4 shows that the bias detected could only explain some of the overall misfit and there might be other sources of misfit as well. Drawing on the think-aloud data, rater 5 repeatedly mentioned that she puts more weight on task completion in comparison to two other criteria. She stated that as long as the speaker performs well on task completion, she would be more lenient with the other two criteria.

Another significantly biased interaction occurred between rater 4 and task completion criteria. Rater 4, who was the most lenient rater among all raters, has a severity measure of -.54. When rating task completion, he rated the speaking performance .81 logits higher than expected. The mean square fit statistic of rater 4 was 1.4, which suggests that the bias found could explain some of the overall misfit. The last significantly biased interaction was found between rater 1 and style. Rater 1 was lenient on rating the style criterion by 1.15 logits. The fit statistics of .9 (which is lower than 1) shows that the rater maintained the identified bias pattern across all examinees.

Relevance to PIE and Second Language Learning

The findings revealed that there were significant biases between the raters and the criteria. One of the factors stated by the raters was about how they see the task and what outcomes they want the students to gain by exercising this task. For instance, rater 3 who was an expert in teaching listening/speaking and working with the same rating criteria mentioned that she would intentionally be more lenient on collaboration and more severe on task completion and for the last category, style, she thought that she would automatically give a high grade if the other two categories earned good score.

Another finding of the study was about the 1-5 scale used in the rating criteria. It was revealed that raters were not using the full range of scale; therefore, there is no need for a 5-point scale for this task. I would suggest changing the scale into a 3-point scale instead.

One of the main recommendations based on the findings of this study is to consider including potential sources of bias in the raters' training session. This may make the rater's aware of what might cause them to be inconsistent in awarding scores. Based on the qualitative data collected in this study, the raters mentioned that the question about the possibility of valuing one criterion over the other made them think and they could all think of a factor contributing to their possible over-emphasis on a certain criteria. For instance, rater 3 repeatedly mentioned her experience as an expert in teaching writing as a possible factor for her to be more severe on task completion criterion more than others. She suggested adding notes for the raters to give credit for the descriptor, "Listens to and responds in order to create cohesion/flow in conversation", only if it makes sense in relation to the other participants' comments.

Another source of bias that was revealed in the think-aloud sessions was the bias towards L1. Rater 5 and rater 2 expressed their bias towards the examinees' L1. For instance, they mentioned that since the Arabic and Portuguese speakers would talk more, they are more inclined to be lenient on their performance on task completion. However, this bias was not proved via FACETS.

Another factor that was brought up by the raters was their familiarity with the examinees and their accents. Rater 2 and 5 stated that they knew many of the examinees. This should be taken into consideration when assigning raters to examinees, since it might produce bias towards certain examinees.

The overall impression of the raters was that the rubric was clear and easy to understand. However, they mentioned having difficulty quantifying the features on the style criterion. They suggested listening more than once helped improving this problem. A suggestion would be to turn the rating sessions for the achievement tests into a focus group and make sure that everyone has had a chance to follow all the features on the rubric. In general, qualitative data, such as think-alouds or focus groups could help inform the training sessions and reduce bias.

References

Barkaoui, K. (2007). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Ed.) *The companion to language assessment* (pp. 1-22). Cambridge, England: Wiley.

Barkaoui, K. (2013). Multifaceted Rasch Analysis for Test Evaluation. *The Companion to Language Assessment*.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*, 3-31. doi:10.1191/0265532202lt218oa

Linacre, J. M. (2014). Facets (3.71.4) [Computer software]. Retrieved from http://www.winsteps.com/facets.htm

McNamara, T. (1996). *Measuring second language performance*. New York, NY: Addison Wesley Longman.

Nakatsuhara, F. (2004). An investigation into conversational styles in paired speaking tests. *Unpublished MA thesis, the University of Essex.*

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465–493.

Taylor, L. & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts Language. Testing, *26*(3), 325-339.