International Journal on Studies in English Language and Literature (IJSELL)
Volume 4, Issue 10, October 2016, PP 66-72
ISSN 2347-3126 (Print) & ISSN 2347-3134 (Online)
http://dx.doi.org/10.20431/2347-3134.0410011
www.arcjournals.org

Comparing Peer and Professor Application of Rating Scales to Speech Performance

Dr. Shaun Justin Manning

Hankuk University of Foreign Studies, South Korea

Abstract: *Universities around the world have been adding English conversation classes to improve oral communicative competence among graduates. This has led to the problems of how to give useful formative feedback and how to reliably and validly assess learners. Peer assessment may be a potential solution. This study sheds light on how students use a rating scale to evaluate other students' speech. It took place in a South Korean university setting. Of interest were the following questions. (1) Can undergraduate university students assess their peers in a way that is comparable to how their instructors rate them? (2) How useful is the feedback from the rating to the recipient and to the provider? To answer these questions, 45 undergraduates recorded a one-minute answer to a prompt similar to an iBT TOEFL prompt. Each student then assessed 15 randomly selected recordings, whichwere also rated by four professors from the same university. Multi-faceted Rasch Measurement (MFRM) techniques were used to examine the ratings. Similar to findings in studies on writing, peer-raters were generally more lenient than instructors. Students reported that the activity helped them think like the teacher. They also made useful contributions to developing the scoring rubric.*

Keywords: *peer assessment, speaking tasks, EFL, Rasch analysis, classroom assessment*.

1. INTRODUCTION

There is a growing demand for school systems and standardized tests to reflect the ability of test candidates' production of spoken English. In recent years, we have seen the introduction of the iBT TOEFL and the introduction of 'add-on' speaking and writing portions of the popular TOEIC test. In South Korea some 'special' high-schools, most universities, and major companies either require English essays or interviews for admission or have special-case entrance through the use of English. However, the teaching and assessment of English productive skills in the South Korean school system remains behind the curve as students, parents and other stakeholders do not trust these types of measures. What is greatly needed is a common, trusted approach to measuring student performance that can be applied to any learner in any situation and be understood by others (Bond & Fox, 2001).

In addition, if schools move to performance assessment, this places great demand on teachers and class time, as these types of assessment typically require a great deal of time and energy. One way around this is to use peer assessment or self-assessment in the process. Neither of these options will be useful, however if the students assessments are unfairly biased.

Multi-faceted Rasch measurement (Linacre, 1989) provides a solution, by expressing rating scores in terms of a common unit and by using information on the performance of the raters to provide 'fair' averages, that is scores that have statistically eliminated biased raters. It cannot be used for only one rater (i.e. a single teacher) so this paper argues for the inclusion of peer raters in this fair average. Including peer raters will improve students' metacognitive awareness of their performance and provide transparency of the teacher's scores. This paper attempts to show how this tool can be used in a classroom setting to provided fair and consistent scores to students.

2. LITERATURE REVIEW

Rating Bias

If we assume that teachers (or trained raters) are both honest and competent, we must still account for discrepancies among them when rating the same performance. In other words each rater will rate in a

different manner. Some raters will rate some aspects (facets) of a performance more severely than other aspects. Other raters will ignore the rating rubric and rate in their own way. For example: It has been found that *grammar* is often the most severely rated aspect on a performance test *even when the test was a communicative test*. (e.g. McNamara, 1996, reporting on the Occupational English Test). Moreover, some raters will simply be more severe than others. And finally, some scores will be differentially more difficult to attain than others, i.e. the degree of difficulty to go from level 1 to level 2 may not be the same as that required to go from level 7 to level 8.

The usual answer to the problem of rater bias is 'train the raters.' This does not work. For the most part, these biases are very stable and resistant to training. There is considerable evidence that training of raters will make them more internally reliable but will not eliminate bias (Lumley and McNamara,1995; Satio, 2008). Lumley (2002, 2005) also noted that there is a tension between a rater's reaction to a paper (the studies were on assessing writing) and his/her attempts to apply the rating scale. McNamara (1996) questioned even trying to eliminate rater bias, preferring to identify and account for it in the final scores by either: (a) multiple marking and averaging the scores, or (b) through the use of Multi-faceted Rasch measurement, a subject to which we now turn.

Multi-faceted Rasch Measurement (MFRM)

A number of researchers have investigated what raters do when they rate using multi-faceted Rasch measurement (MFRM) (Linacre, 1989, 1998). Each rater looks at different aspects of a test situation differently. Aspects of the rating situation are called 'facets', and they include:

1. Candidate ability
2. Item / question difficulty: This can be further divided in the case of speaking tests to include: Pronunciation, grammar, topic development, communicative effectiveness, intonation, etc.
3. Rater severity

MFRM allows each facet to be measured in a interval scale using a measurement called a 'log odd unit' (logit). Facets are examined independent of each other but are compared to all other facets in the testing situation. We can now examine the interaction between the candidate, the item, and the rater (using a common scale – the logit scale). We can even look at the interaction of the rater *at a given level on the scale* for an item with the candidate.

Log Odd Units and Probability

A log odd unit (logit) is a standard measure for that data set. It is calculated by comparing the logarithmic probability (calculated by the FACETS program, Linacre, 2009)of the candidate getting this score from this rater by incorporating information on how the rater rated other candidates with respect to other raters – was the rater stricter, easier, more biased against grammar errors, etc. than other raters were. To do this, there must be some overlap among the raters so their scoring behaviour can be compared.

When using logits to describe item difficulty, a useful way of thinking of logits is that when an item and a candidate are at equal points on the scale (the test is at the student's level) the student has a 50% chance of success. If the student's ability is 2 logits greater than the test item's difficulty, s/he has an 88% chance of success.

When using logits to describe rater severity, we can say that if a candidate were scored by a rater 2 logits lower(more strict) than an average rater, the student will have about a 12% chance of getting the appropriate rating for the performance. So we can use logits to adjust scores, as opposed the highly unfair practice of using raw scores for candidates. This can also be used to train raters or to make decisions as to whether or not to employ / continue to employ a rater. For this study, I wish to use this information to provide feedback to students about how scores are assigned and what aspects of language they need to work on.

Previous studies using MFRM

Schaeffer (2008) looked at the rating of essays in Japan. He asked 40 NES to rate 40 Japanese students' essays using a 6-item scale: Content, Organization, Style and Quality of Expression,

Language Use, Mechanics and Fluency. The essays were graded on a 6-point scale for each item. He found several subgroups, among them. First, some NES would rate Content/ Organization severely and then compensate by marking Language Use/ Mechanics leniently. Also, some NES marked the higher level students more severely and the lower level students more leniently than expected.

Matsuno (2009) looked at self, peer and teacher assessments of writing (91 Ss and 4Ts) in Japan. What she found with respect to self-assessment was a tendency for most high level students to score themselves low and that S.A. was "idiosyncratic" and "of limited utility as a part of formal assessment." With respect to peer-assessment she found that peers were the most lenient raters, peers were internally consistent, and peers exhibited the tendency to rate high achievers lower, and low achievers higher was independent of their own ability. With respect to teacher assessment, she found that teachers were internally consistent, and that each teacher had their own unique bias pattern. Finally, in her examination of the items on the scale, she found that grammar was rated severely by all groups, and that spelling was rated most leniently. Her conclusion was that peer-assessment could be part of a writing program if MFRM were used to help Ss identify their biases. This current study attempted to replicate and expand her findings to the measurement of speaking.

3. METHODS

The research questions for this study were:

1. Can undergraduate university students use the instructor's rating scale to reliably assess their peers in a way that is comparable to how their instructors rate them?
2. How useful is the feedback from using the rating?

To answer them, a mixed methods approach was undertaken. The first question was addressed by a statistical analysis of the peer analysis data. Peer ratings were compared to professors' ratings. The second was addressed by a post-study interview.

Participants

The participants in the study were 42 students studying a 2nd year class called 'Advanced Conversation' which was taught by the researcher. There were 18 female and 26 male students. Four maleprofessors from the same university also rated the samples. Each had at least 10 years teaching in this context.

Data collection (Voice recording assignment)

Recording homework.

This study was intended to shed light on an existing classroom practice, so a regular assignment was chosen for analysis. This was a recording assignment. Each student submitted a one-minute mp3 recording of their spoken answer to a given question, patterned on the iBT TOEFL's independent speaking prompts and relating directly to the topic of conversation discussed that lesson. Once the students had recorded their answer, they had to email it to the teacher who then uploaded the recordings to a password protected site on the Internet.

Rating procedure

Students were given a 20-minute training session using a previous year's assignment as sample. They then rated the recordings as a homework assignment using the scale below which was adapted from McNamara (1996) with a slash / between the 3 and 4 points on the scale indicating the 'Pass – Fail' line.

Levels were assigned as follows:

1 = weak, 2 = developing, 3= below average, 4 = above average, 5 = good, 6 = very good.

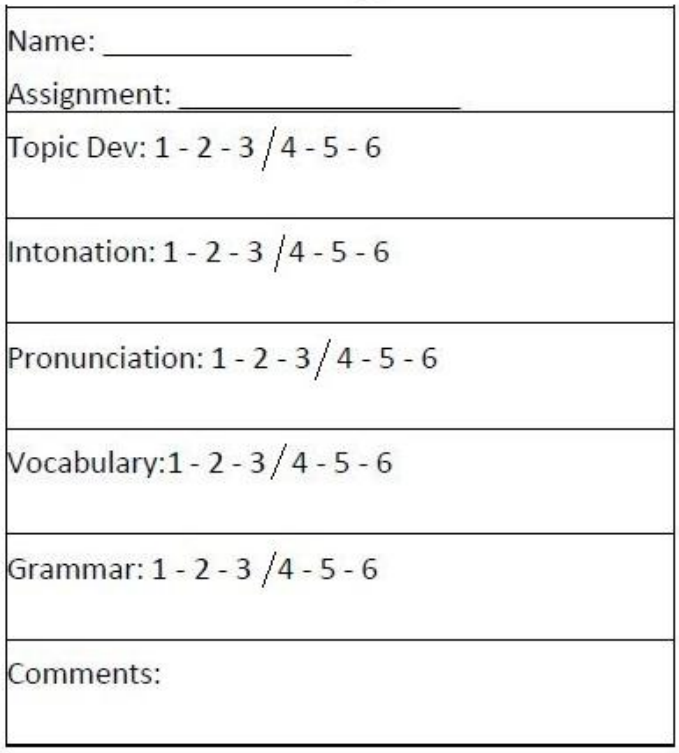

Figure1. *Rating sheet filled out by students*

Students listened to each recording two times and then marked their ratings on an excel file. When the rating was completed, the file was emailed to the instructor. The instructor then complied the data and used the statistical program *Facets* (Linacre, 2009) to do the multi-faceted Raschanalysis.

Interviews

After the ratings were sent to the instructor, seven students carried out semi-structured interviews to determine what students thought of the procedure, and how useful they felt it was. The interviews were carried out by a research assistant and conducted in Korean, so the students could express themselves fully. The research assistant transcribed and translated the interviews, which were then analyzed discursively, looking for emergent themes and possible intervening variables that may be acting on the quantitative data.

4. RESULTS AND DISCUSSION

The facets map below (See Figure 2, below) shows the results of both the peer and instructor ratings.

There are six columns in the figure. The left-hand column shows the logit scale. If readers imagine a horizontal line across the chart at 0, anything above this line is 'lenient' or 'high' or 'better performance' whereas anything below is 'strict', 'low', or 'weaker performance'. The chart is scaled (10 dashes per logit) so we can calculate the differences among facets from simply looking at the chart.

The second column shows how the raters spread out. When reading the chart, each *M* indicates 2 male students and each *F* is 2 female student raters, each *P* represents 2 professors. The important information from this chart is that the range of leniency is very wide. The strictest raters were at -0.8 logits, whereas the most lenient were at +2.8 logits. This means that if a student were scored by the latter rater they would have well over 90% probability of achieving their score, while a less than 40% chance with the strict rater. This was a key question in the interview (see below). The professors are 0.8 logits separated as well, meaning the more lenient ones will be preferred by the students! This implies, that the 'fair' average computed by FACETS should be used rather than the raw scores of the professors for official grading.

The third column shows that student recording 'a' was fully 2 logits worse than the best students (d and j). The fifth column shows that Topic Development was the most leniently graded facet, whereas

Intonation was the most severely graded facet.

The fourth column shows that the native language of the rater was not a factor in the overall ratings. The 'classifier' column shows that Native speakers of Korean (NSK), native speakers of English (NSE), and the one native speaker of Spanish (NSO) used the rating system equally severely overall. NSK'sfair average on all facets was 4.33, NSE's fair average was 4.30, and NSO was also 4.33. Infit mean squares were all within acceptable limits of 0.5 to 1.5 (cf. Linacre, 1989, 1998). The implication is that with training, non-native speakers can rate the speech of their peers at an overall level similar to native speakers. However there is the issue of the wide variability among the students.

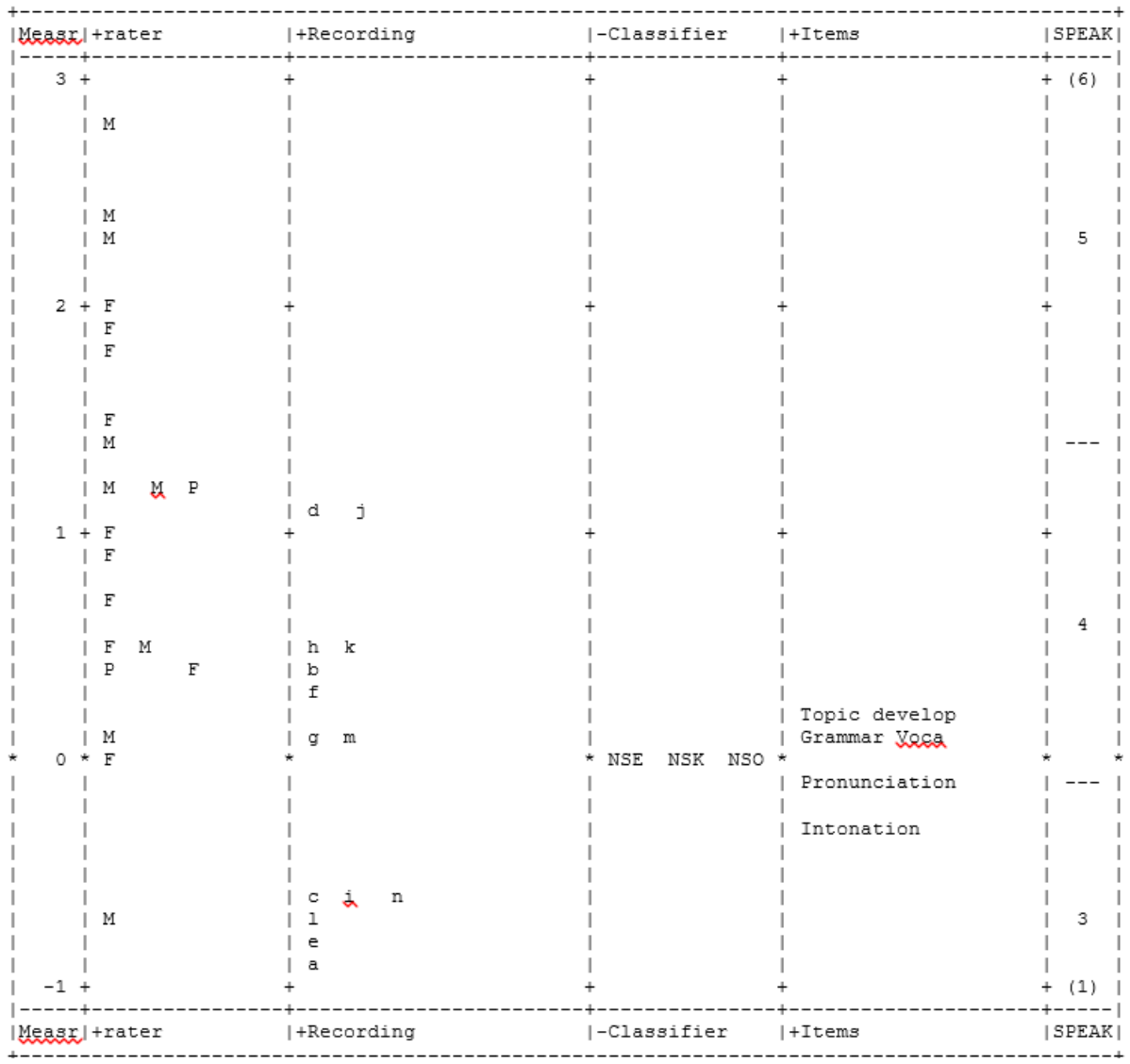

Figure2. *Facets map for student ratings*

One last issue is the rating scale itself needs to be shown to 6 equidistant parts. Figure 3, below, shows that this was true for native speakers but not for the Korean-speaking students. The native speakers were able to do this. Each 'step' from 1-6 was at least 1.4 and not more than 5.0 logits (Linacre, 1989). However, the students were not able to separate all the steps. Steps 4 and 5and 5 and 6 are too close to each other (see Fig 3, below).

In Figure 3, the Peer Rating Scale has a gap of only 0.64 between 4 and 5. This means, according to the students it is easy to go from a 4 to a 5. Similarly, the gap between 5 and 6 is 1.12 logits, also implying an easy step. Conversely, the gap of 2.05 logits between 2 and 3 and 3.18 logits from below average (3) to average (4) implies that students were not willing to give poor scores to their peers, so they, as raters made it easier to get high grades and more difficult to get lower ones.

Interview data bears this out. When asked about how they scored, most students said they started at the middle, where the slash mark was, and went up or down, they did not start at the bottom and go up,

which was the procedure followed by the professors. The result is more student performances receiving higher scores from the student raters.

Peer Rating Scale		Professors' Rating Scale	
Category	**Threshold (SE) (logits)**	**Category**	**Threshold (SE) (logits)**
1 'Weak'	-∞	1 'Weak'	-∞
2 'Developing'	-4.34 (.16)	2 'Developing'	-3.32 (.28)
3 'Below average'	-2.29 (.14)	3 'Below average'	-1.85 (.11)
4 'Average'	0.89 (.08)	4 'Average'	-.11 (.07)
5 'Good'	1.53 (.06)	5 'Good'	1.44 (.06)
6 'Very good'	2.65 (.10)	6 'Very good'	3.84 (.08)

Figure3. *Comparing student rating scales for homework recordings vs. discussion tests*

Another way to look at this is to compare the scores a student of a specific ability would have received from a peer vs. what a professor would have given as in Figure 4, below.

Student ability (logits)	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
Peer score (rating scale)	1	2	3	3	4	4	5	6	6	6	6
Professor Score (rating scale)	1	1	2	3	3	3	4	5	5	6	6

Figure4. *Comparison of scores students of a specific logit ability receive*

Figure 4 compares the scores that students would have received if they were rated by professors or other students. It shows that each score from 1 – 6 starts at a lower score for the student raters than for the professor raters. For example, a student rater will give a performance of 1 logit ability a score of 5, while a professor will only give a 4. Likewise a student rater would give a performance of 2 logit ability a score of 6, while a professor would give a 5. This very low threshold for 6 reflects one common complaint of peer review, the students are too lenient.

When asked about this in interviews, the students reported that they felt they should not give a 1 or 2 to their classmates because they would feel bad. Even though the rating was anonymous, the students could not bring themselves to give a low score. Two interviewees suggested changing the rating scale itself. They wanted it to be out of 20 or 25, and in the case of the latter, only use numbers 20-25 to give the 6-level scale. That way they would not feel bad about assigning a 20.

They also suggested removing the slash from the scoring sheet. McNamara (1996) had included it to push raters to make a decision about passing or failing, but for classroom purposes, it was thought to be unhelpful. The peer raters did not like failing their peers, as shown in Figure 4. To fail (get a 3) a performance needed to be -2 logits in the students' eyes, but only 0 logits in those of the professors.

However, the interviewees also said that the activity was helpful because it clarified the professor's scoring criteria and what was expected of them. They also said that they would focus on parts where they did poorly next time, because they realized how they were being graded.

To sum up the results, student raters saw the peer rating activity as helpful to their language development, but they did not apply the rating scale as the professors did. This mismatch became the topic of a later conversation class because students worried about the strictness of the scores and whether they could achieve a top score. For the instructor, the activity was enlightening as it informed him about the need for clarity in grading and the need to give students a voice.

5. IMPLICATIONS AND CONCLUSIONS

The implication of this small study is that using Rasch techniques can inform students about their strengths and weaknesses as students, can help students learn how to think like their teacher, and help them change their approach to learning. It helped the teacher learn about his students' worries and concerns and showed him how to better communicate with his students about their scores and how they are derived.

ACKNOWLEDGEMENTS

This paper was supported by the Hankuk University of Foreign Studies Research Fund for 2016.

REFERENCES

Bond, T.G. & Fox, C.M. (2001) *Applying the Rasch Model: Fundamental measurement in the human sciences*New Jersey: Lawrence Earlbaum Associates

Linacre, J. M. (1989, March). Rasch models from objectivity: A generalization. Paper presented at the International Objective Measurement Workshop, Berkeley, CA.

Linacre, J. M. (1997). Guidelines for rating scales. MESA Research Note #2. Retrieved January 8, 2003, from http://www.rasch.org/rn2.htm

Linacre, J. M. (1998). Rating, judges, and fairness. *Rasch Measurement Transactions, 12*(2), 630–631. Retrieved September 9, 2002, from http://www.rasch.org/rmt/rmt122f.htm

Linacre, J. M. (2009). FACETS (Version 3.65.0) [Computer Software]. Chicago,IL: MESA Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lumley, T. (2005). Assessing second language writing: The rater's perspective. Frankfurt am Main: Peter Lang.

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

Matsuno, S. (2009) Self-, Peer- and Teacher-assessments in Japanese university EFL writing classrooms. *Language Testing 26* (1), 75-100

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

Saito, H. (2008) EFL Classroom peer assessment: Training effects on rating and commenting. *Language Testing*, *(25)*4, 553-581

Schaefer, E. (2008) Rater bias patterns in an EFL writing assessment. *Language Testing*, *(25)*4, 465-493

AUTHOR'S BIOGRAPHY

Shaun Justin Manning, holds a PhD in Applied Linguistics from Victoria University of Wellington, NZ. His research interests are: instructed SLA, task-based learning, task design and implementation, and task-based assessment. He teaches undergraduate English proficiency and post-graduate TESOL classes at Hankuk University of Foreign Studies.