Validity of Peer Assessment of Speech Performance

Makoto FUKAZAWA
Ibaraki Prefectural Takezono High School

Abstract

The main purpose of this study is to examine how valid peer assessment can be in a high school setting in Japan. One hundred fifty seven students and five teachers participated in this study. The observed scores of the peer and teacher assessments were analyzed using Multi-faceted Rasch Measurement. According to the Rasch analysis, 94% of students did not show misfitting behavior, which suggested acceptable validity of peer assessment (Myford & Wolfe, 2003). Peer assessment was also examined at a class level, and this also fitted into the model. In addition, characteristics of misfit raters were investigated in terms of unexpected responses and English proficiency level, but no specific characteristics in common were found. However, some overall tendencies for peer assessment were observed. The average scores of peer assessment were higher than those of teacher assessment, whereas the standard deviations of peer assessment were smaller than those of teacher assessment. These results indicate that peer assessment of Japanese high school students has reasonable validity in a certain context, although student raters tend to be more lenient compared to teacher raters, and they tend to avoid extreme scores.

1. Introduction

Improvement in practical communication skills has been the focus of secondary school English education in Japan. Its importance has been emphasized and practical action has been taken through the “Action Plan to Cultivate ‘Japanese with English Abilities’” and “Super English Language High School (SELHi)” by the Ministry of Education, Culture, Sports, Science, & Technology (MEXT). In addition, the Course of Study for Upper Secondary School (MEXT, 2009), emphasizes the integration of four English skills—reading, writing, listening, and speaking—and will be put into practice in 2013.

Through these practices above, students have had more opportunities to improve their communication skills. Speaking skill is, of course, not an exception. However, as Akiyama (2000) suggests, speaking assessment is not focused on as much as its teaching methods. The reason is that speaking assessment is considered to be one of the most difficult fields of assessment (Baba et al., 1997) because of its subjectivity (McNamara, 2000) and the large number of students in language classrooms in Japan.

In order to solve or at least reduce some problems of speaking assessment, the use of peer
assessment (PA) is considered. PA can add another perspective in a language classroom, which makes speaking assessment more objective. It can also help to reduce teachers' workloads to make speaking assessment more practical. However, the validity of PA has not been firmly confirmed as a part of formal assessment. In this research, therefore, the validity of PA was examined using Multi-faceted Rasch Measurement (MFRM).

2. Literature Review

PA is defined as "any assessments that require students to judge the language or language performance of one or more other students (or peers)" (Brown, 1998, p.54). The advantage of PA is its feasibility, because it does not require much time and effort (Brown). PA also motivates students (Orsmond, Merry, & Reiling, 1996), making learning objectives clearer. Brown, on the other hand, points out that PA is a relatively subjective method of assessment and it can be unreliable in high-stakes situations, such as final examinations.

Therefore, if PA is considered to be used to assess speaking, especially as a part of summative or formal assessment, the validity of PA must be examined. Miller and Ng (1996) examined the validity of PA with 41 college students in a speaking class, comparing the results of PA and those of teacher assessment (TA). There were relatively high correlations between PA and TA \((r = .68- .80)\) and they concluded that PA was comparable to TA. They also suggested students with high English proficiency could be more reliable peer raters. Stefani (1994) compared TA, PA, and self-assessment (SA). When comparing correlations between these values, PA and TA had the highest correlation \((r = .89)\) and concluded PA, unlike SA, can be as valid as TA. Stefani also found a tendency for the average score of PA to be higher than that of TA and that the standard deviation of PA was smaller than that of TA.

There has only been a little research done on PA using MFRM. Nakamura (2002) examined the validity of PA in an oral presentation class using the Rasch Model. Five student raters were randomly chosen from the class and four out of the five raters fitted into the model, which indicated that student raters functioned effectively. Based on this result, Nakamura concluded that students could be reasonably reliable raters. Matsuno (2009) compared the validity of TA, PA and SA in a writing class. The result suggested that most of the students had internal consistency as peer raters and PA was the least biased when compared to TA and SA.

On the other hand, other research showed negative results for PA. Orsmond, Merry, and Reiling (1996) analyzed their results of PA using regression analysis. The results showed a high correlation \((r = .89)\) between PA and TA, but there was a significant difference in the standard deviations between them. As a result, they questioned the validity of PA. Freeman (1995) also conducted research with 210 college students and compared results using regression analysis. Only moderate correlation between PA and TA was observed, and there was a tendency for students to evaluate excellent performances lower, and poor performances higher.
As the previous studies on PA show, the results contradict each other. Therefore, the main purpose of this study is to examine the validity of PA of students' speech performance, so as to discover the probability of using PA in high school English education in Japan. In order to achieve this goal, the following research questions were set.

RQ1: To what extent PA of speech performance is valid at an individual level?
RQ2: To what extent PA of speech performance is valid at a class level?
RQ3: Are there any common characteristics or tendencies among misfit raters or in classes that do not have acceptable validity?

3. Method

3.1 Participants

The participants in this study were 157 Japanese public high school students (80 first year students and 77 second year students). Among these students, five students have foreign nationalities, but none of them come from English speaking countries. This experiment was conducted in a high school where the researcher works and the school is considered as an academically strong high school, which emphasizes science and English education. Four Japanese teachers of English (JTEs) and one assistant language teacher (ALT) also participated in this study. Each of JTEs has more than 10 years of high school English teaching experience, and the ALT has taught English in Japan for three years.

3.2 Materials

A proficiency test, rater training video, assessment criteria, and evaluation sheet were made and used as materials for this research. The proficiency test consisted of three parts—listening comprehension, grammar, and vocabulary sections—and contained 50 questions in total. Freeman (1995) suggests the use of a video for rater training for reliable PA, so a rater training video was made to show students and teachers the six levels of speeches based on the 6-point Likert scale. The rating scale was designed for Japanese high school students based on Common European Framework (Council of Europe, 2001). The assessment criteria were pronunciation, accuracy, delivery, fluency, and overall evaluation. According to these the rating scale and criteria, an evaluation sheet was made.

3.3 Procedure

This research was conducted using four classes; two classes (A and B) from second year and two classes (C and D) from first year. The students were told that the results of the PA were included as a part of their grades as well as the teachers' evaluations. First of all, the students had a 30 minute proficiency test ($\alpha = .78$). Then, the students had rater training with the training video using one period. Watching the video, students saw sample speeches of each level and then they
had assessment practice three times. Rater training for the teachers had been done in advance, following the same procedure. After the training session, students made speeches using three periods (13-14 speeches in each period). After each speech, one minute of PA followed. All the students in the classes, except for the speaker, evaluated the speech as PA. Three or four teachers also evaluated the students' speeches at the same time.

4. Results and Discussion

4.1 Overview

The results of the PA and TA were analyzed using FACETS (version 3.64.0), which is a computer program for MFRM. In this study, only the results of the overall assessment were analyzed of the five assessment criteria because the focus of this research is the validity of PA in general. From 157 participants, eligible data was available for 145 speakers and 117 raters. The amount of eligible data for speakers was less because some students’ speeches were too short to evaluate. The amount of eligible data for raters also was even less because some students could not evaluate all speeches.

Figure 1 shows the overview of the results of the Rasch analysis done. In the FACETS valuable map, the first column from the left is a logit scale. Logit is a scale used for MFRM and is a true interval scale, unlike raw test scores. On this scale, zero indicates the average. The second column shows item difficulty, and in the case of this study, item difficulty is the students' ability to make speeches. Each asterisk (*) represents two speakers and a dot (.) represents one. If the marks appear in the upper part of the column, it means the speakers have higher ability, and marks in the lower part mean the speakers have poor ability. The third column displays rater severity. The more severe raters appear in the higher part of the column, and the more lenient raters appear in the lower part. In this column, each asterisk represents one rater, T represents a teacher rater, and M indicates a misfit rater, who does not fit in with the model. The variables, or raters in this case, range from -3 to 3 logit in general, although logit can be theoretically infinite. Considering the general range, one very strict rater and three very lenient raters are seen in this column. The fourth column also shows rater severity as a class. The letters in the column indicate classes, and T indicates the teachers as a group. There are not many differences in severity among classes, though class D is slightly more severe than the other classes and class B is slightly more lenient. The last column indicates the 6-point scale according to the assessment criteria (C2 is the highest, and A1 is the lowest). Taking a look at both the second and the fifth columns, most speakers fall into categories B1, B2, and C1, only a few students belong to categories A2 and C2, and none of the students fall into the A1 level.
Table 1 Descriptive Statistics of PA and TA

<table>
<thead>
<tr>
<th>Class</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.97</td>
<td>0.77</td>
<td>3.78</td>
<td>0.93</td>
</tr>
<tr>
<td>B</td>
<td>3.91</td>
<td>0.79</td>
<td>3.51</td>
<td>0.94</td>
</tr>
<tr>
<td>C</td>
<td>3.79</td>
<td>0.78</td>
<td>3.59</td>
<td>0.80</td>
</tr>
<tr>
<td>D</td>
<td>3.94</td>
<td>0.77</td>
<td>3.92</td>
<td>0.81</td>
</tr>
</tbody>
</table>

*Note. n = 30 (Class A), n = 28 (Class B), n = 30 (Class C), n = 29 (Class D). M = mean. SD = standard deviation.*
Table 1 shows the descriptive statistics of PA and TA. According to the data, two tendencies of PA can be observed. One tendency is that the mean scores of PA were consistently higher than those of TA. This indicates that the student raters tended to give more generous scores compared to the teachers. The second tendency is the standard deviations of PA were consistently lower than TA’s. This shows teacher raters used a wider range of scores, whereas peer raters tended to avoid extreme scores such as A1 and C2. As Freeman (1995) stated, the students might give lower scores for good speeches and better scores for poor ones. These two tendencies—PA’s higher mean scores and lower standard deviations when compared to TA’s—support the results of Stefani (1994).

4.2 Research Question 1

Research question 1 asks to what extent Japanese high school students’ PA of speech performance is valid at an individual level. This question was answered by the infit mean square variable of each student rater. This is because the infit mean square shows the degree to which observed scores match the expected scores by MFRM, which indicates how effectively raters can assess ratees in terms of their level of performance (Myford & Wolfe, 2003). The expected value of the infit mean square is 1.00. As values tend toward 1.00, observed ratings are closer to expected ratings from the Rasch Model. The acceptable range of the infit mean square valuable varies depending on assessment context. Myford and Wolfe suggested an upper-control limit of 1.20 and a lower-control limit of 0.80 for a high-stakes test, while 2.00 for an upper-control limit and 0.50 for a lower-control limit in the case of low-stakes decision making. McNamara (1996) suggested that the range should be the mean ± twice the standard deviation of the mean square statistic, and that this is more precise in the case of n sizes of 30 or more. This study takes McNamara’s standard as the acceptable range because the sample size is over 30 and the results can be more precise.

Table 2 Rater Measurement Report for Individuals

<table>
<thead>
<tr>
<th></th>
<th>Infit MnSq</th>
<th>ZStd</th>
<th>Outfit MnSq</th>
<th>ZStd</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0.98</td>
<td>-0.20</td>
<td>1.05</td>
<td>-0.10</td>
</tr>
<tr>
<td>SD</td>
<td>0.38</td>
<td>1.50</td>
<td>0.82</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Note. N = 122. Infit MnSq = infit mean square. ZStd = z standard. Outfit MnSq = outfit mean square. M = mean. SD = standard deviation.

Table 2 shows the mean and the standard deviation of infit mean square values of raters. In the case of this study, where the mean score of the infit mean square was 0.98 and the standard deviation was 0.38, the acceptable range was 0.36 to 1.74 (0.98 ± [2 × 0.38]). As is shown in Figure 1, seven raters did not fit the model. The infit mean square values of the misfit raters were 2.36 (Rater D15), 2.15 (R C29), 2.14 (R A28), 2.04 (R B03), 1.86 (R B25), 1.84 (R A12), and 1.81 (R B07). In general, misfit raters and misfit speakers (eight speakers did not fit the model in
this study) should be eliminated for the second Rasch analysis in order to obtain more accurate data, but eliminating these raters and speakers is pedagogically unrealistic. Therefore, the results of the initial analysis were examined in this study. Seven misfit raters out of 117 translates to 6.00% of all student raters. In other words, 94.00% of the students were internally consistent raters. Regarding students as individual raters, it is reasonable to claim that most of the students have acceptable validity as raters, which supports the result of Nakamura (2002).

4.3 Research Question 2

Research question 2 asks to what extent Japanese high school students’ PA of speech performance is valid at the class level. This question is asked because all students in a class, except for the speaker, evaluated speeches and it is considered to be one of the practical types of PA. Research question 2 was mainly examined by infit mean square values, but another aspect, severity, was also considered.

Table 3 Rater Measurement Report for Classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Fair Mean Average</th>
<th>Measure</th>
<th>Infit MnSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.89</td>
<td>0.11</td>
<td>1.06</td>
</tr>
<tr>
<td>B</td>
<td>4.03</td>
<td>-0.41</td>
<td>1.04</td>
</tr>
<tr>
<td>C</td>
<td>3.95</td>
<td>-0.12</td>
<td>0.87</td>
</tr>
<tr>
<td>D</td>
<td>3.85</td>
<td>0.26</td>
<td>0.94</td>
</tr>
<tr>
<td>T</td>
<td>3.88</td>
<td>0.16</td>
<td>1.16</td>
</tr>
<tr>
<td>M</td>
<td>3.92</td>
<td>0.00</td>
<td>1.01</td>
</tr>
<tr>
<td>SD</td>
<td>0.06</td>
<td>0.24</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Note. *n* = 30 (Class A), *n* = 28 (Class B), *n* = 30 (Class C), *n* = 29 (Class D), *n* = 5 (Teacher raters). Infit MnSq = infit mean square. *M* = mean. *SD* = standard deviation.

According to Table 3, the acceptable range of the infit mean square was 0.81 to 1.21 (1.01 ± [2 × 0.10]). The infit mean squares of each class ranged from 0.87 (class C) to 1.06 (class A), so the infit mean square values of all classes fitted into the acceptable range. Compared with the infit mean square value of the teacher group, the values of each class tended toward 1.00, which means their scores were even closer than those of the teachers to expected scores from the model. The results indicate that every class maintained sufficient validity for assessing speeches.

The rater severity of each class was also taken into consideration. The measure values of Table 3 show the severity of each class. The most severe class was D, with a logit score of 0.26, and the most lenient class was B, with a logit score of -0.41. As also seen in Figure 1, all of the classes are located near the average in the column, and no remarkable differences in severity can be seen in each group of raters, including the group of teachers. Each class had a sufficient level of internal consistency, and their severity levels were similar to each other. These reasons support that PA, as a class, has reasonable validity which is comparable to TA.
4.4 Research Question 3

In order to find out any characteristics or tendencies misfit raters might have in common, their unexpected responses provided by FACETS and their English proficiency levels were investigated. First, misfit raters' unexpected responses were closely looked at. Table 4 shows the number of unexpected responses that misfit raters made and the severity of those responses compared to expected responses provided by the Rasch Model. The number of unexpected responses for each misfit rater varied. For example, rater D15 had five unexpected responses while B03 had only one. Other raters who fitted the model had unexpected responses in a similar manner, so no remarkable difference was observed in terms of the pattern of unexpected responses. Among the 21 unexpected answers of misfit raters, 11 responses were too severe and the other 10 were too lenient. The numbers of 'too severe' and 'too lenient' responses were almost the same, so no obvious tendency was observed in terms of severity, either. The only observation worth mentioning from Table 4 is that more misfit raters belonged to classes A and B compared to classes C and D. As is mentioned above, classes A and B were second year classes and the other two classes belonged to the first year.

<table>
<thead>
<tr>
<th>Misfit Rater</th>
<th>Unexpected Response</th>
<th>Too Severe</th>
<th>Too Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A12</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A28</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B03</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B07</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>B25</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C29</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>D15</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>21</strong></td>
<td><strong>11</strong></td>
<td><strong>10</strong></td>
</tr>
</tbody>
</table>

In order to find out the reason, the fit statistics of the speakers were examined. The mean score of speakers' infit mean square value was 1.00 and the standard deviation was 0.29. Using the same formula, the acceptable range was 0.42 to 1.58 (1.00 ± [2 × 0.29]). According to the limits of this range, there were eight speakers who did not fit in with the model. Their infit mean square values were 1.87 (Speaker B11), 1.78 (S A06), 1.69 (S B18), 1.67 (S B01), 1.65 (S D21), 1.60 (S A16), 1.59 (S B02), and 0.38 (S C25). Four misfit speakers belonged to class B and two belonged to class A. Apparently, there were more misfit speakers in the second year classes, especially in class B. This might have caused problems when peer raters assessed these speakers, probably resulting in more misfit raters in those classes.

Second, the misfit raters were examined in terms of their English proficiency, because being a reliable rater may require a high level of English proficiency as Miller and Ng (1996) stated. For the examination, all participants were divided into four groups, according to the results of the proficiency test given. Then a one-way ANOVA was performed to check the statistical
differences in proficiency between the four groups. It was found that equal variances were not assumed among the proficiency groups; therefore Brown-Forsythe's correction was made, changing the degrees of freedom and F-values, $F (3, 50.93) = 233.36, p < .001$. Then, the Games-Howell Pairwise Comparison Test was conducted for multi-variable comparisons and the differences between the groups were confirmed ($p < .001^*\text{).}$ There were two misfit raters in the top proficiency group ($n = 30, M = 33.90, SD = 4.44$), two misfit raters in the second group ($n = 32, M = 25.50, SD = 1.37$), two in the third group ($n = 29, M = 21.66, SD = 1.17$), and one in the lowest group ($n = 26, M = 21.66, SD = 1.95$). As the number of the misfit raters in each proficiency group shows, there is no apparent explanation according to the misfit raters' English proficiency, unlike Miller and Ng suggested.

In terms of unexpected responses and English proficiency level, no specific characteristic was observed in misfit raters. Although the classes which had more misfit speakers had more misfit raters, it might be difficult to generalize the results since the sample size is limited.

5. Conclusion

This study examined the validity of PA in a Japanese high school setting. First, the validity of PA as an individual rater was examined, using MFRM. The results showed most of the high school students fitted the Rasch model, which indicated the validity of PA. Second, the validity of PA was examined as a class. Every class fitted the model and the validity of PA as a class was also confirmed. Based on these results, it is reasonable to conclude that PA has sufficient validity for assessing speeches in a Japanese high school. In addition, misfit raters were closely examined in terms of their unexpected responses and English proficiency level, but no significantly common tendency was found.

However, there are three points to keep in mind when PA is put into practice. The first point is that there might be a few unreliable student raters per class. Therefore, it would be better to conduct PA as a whole class or in relatively large groups in order to maintain the validity of PA. Second, there are tendencies for PA to be more lenient than TA, and PA tends to avoid extreme scores, in comparison with TA. When a teacher uses PA as a part of formal assessment, he/she needs to keep these tendencies in mind. Third, it would be safer not to use PA in a high-stakes context, because the acceptable range of the infit mean square values for this study was 0.36 to 1.74, which is considered to be a relatively low-stakes context according to Myford and Wolfe (2003), and it remains to be seen whether PA can be used in a high-stakes setting.

If the validity and feasibility of speaking assessment can be improved through PA, it will facilitate having more speaking activities in an English classroom. It can also support the more active use of integrated communicative activities, which are strongly suggested in the Course of Study for Upper Secondary School (MEXT, 2009) for English classrooms in Japan.
Acknowledgements

This research was partially supported by the 21st grant for research from the Society for Testing English Proficiency [Eiken Kenkyu Josei]. I would like to thank Rie Koizumi, Masatoshi Nagahashi, Tracey Wilson, and the reviewers for giving me valuable advice.

References


