最終プレゼンテーションにおけるピア評価の考察

An Analysis of Peer Assessment for Final Oral Presentations

大　年　順　子
OTOSHI, Junko

岡山大学大学院社会文化科学研究科紀要
第53号　2022年３月　抜刷
Journal of Humanities and Social Sciences
Okayama University Vol.53 2022

An Analysis of Peer Assessment for Final Oral Presentations

OTOSHI,Junko

1. Introduction

Peer assessment has been utilized in college EFL (English as a Foreign Language) classrooms for various activities, and especially for speaking and writing. Students can benefit from peer assessment in ways that assist their oral output skill, and by giving them a sense of audience from sociocultural points of view (Villamil & Guerreno, 2006). In the case of oral presentations, students can become a judge like their teacher while they are part of the audience for their peers' presentations.

Peer assessment is preferable to use as part of formative assessment; providing diagnostic information is more valuable rather than judging the quality of performances (Hamp-Lyons, 1991). In contrast, summative assessment requires a more valid and reliable assessment quality from the raters. This is the biggest reason why peer assessment has not been established as a solid status yet as an alternative assessment despite a large volume of research in the field (e.g., Cheng & Warren 2005; Patri, 2002; Saito, 2008).

The final presentations, which represent an example of summative assessment, usually become one of the most important considerations for grading to the students' output skills in the course. The classroom teacher has to take a serious responsibility alone in making his or her assessment valid and reliable. However, a single teacher's assessment for all individual students' presentation could be problematic (Matsuno, 2017). In reality, it is uncertain to what extent teacher's assessment is valid or reliable. Marking is a subjective activity (Patri, 2002). A single rating for the most important students' performances poses a question; therefore, multiple assessments should be triangulated for a reliable assessment (Orsmond & Merry, 1996). In this regard, peer assessment is expected to be incorporated into summative assessment.

Moreover, the classroom environment also affects the rating consistency in the EFL context. In college English classrooms in Japan, it is common to have between 30 to 40 students in one class, especially in general English education courses. It usually takes multiple sessions in order to have all the students complete their final presentations in such class sizes. It has been questioned how consistent the teacher's assessment is over the multiple sessions.

Despite those concerns about a single teacher's assessment for students' final presentation, teachers are still hesitant to use peer assessment as part of summative assessment. As mentioned earlier, it is certainly due to its reliability. Students are still language learners. They cannot assess in the same way with their teacher (Cheng & Warren, 2005).

Another caution for peer assessment is the friendship effect. As Tanaka (2017) found, personality traits of individual students affect peer assessment as students get to know their peers more in the classroom. Since the final presentations usually take place in the last sessions of the course, students are already familiar with each other. It is not surprising that the personal traits of individual students might influence their rating quality.

However, such concerns are not limited to the students. Non-native English-speaking teachers may also have insecure feelings towards awarding a score to the students' final presentations. Even if the teacher has a strong command of speaking in English in addition to teaching experiences, he or she also has to switch the role from an educator to a judge for the assessment. The teacher's assessment is also considered as subjective and the students might be worried about its reliability.

One possible solution for improving the quality of assessment is rater training referring to rating scales. When a rating criterion is firmly set and has clearly defined achievement levels, raters can enhance the reliability of their assessment (McNamara, 1996). In this respect, raters who need the training include both teacher and the students. Despite the different roles in the classroom and knowledge for key elements of oral performances, both the teacher and students have to assess final presentations referring to the same rating scale. It is, in other words, hypothesized that both the teacher and students can make a consistent and similar summative assessment when they mutually understand the criteria for assessing the final presentations.

2. Peer Assessment for Oral Presentations in the EFL Context

Most of the studies in the EFL context reviewed for this study have shown the positive associations between peer and teacher assessments for oral presentations; incorporating rater training improves these associations. However, the magnitude of the associations and severity of the two assessments vary from study to study. Also, the approaches to rater training are different and situated in the research context.

As an empirical study for examining the effects of rater training for assessing students' oral presentations, Patri's research (2002) has been frequently cited in the target field. The Chinese college students allocated in the experiment group had received rating training by watching

videos of previous students' oral presentations and practiced rating. Although Patri's research showed a statistically significant association between peer and teacher assessments in the experiment group, the research revealed that students tended to give more generous scores to their peers than teacher assessments.

Cheng and Warren (2005) investigated peer assessment implementing rater training sessions throughout an academic English course for both teachers and students. They found that 51 first-year college students' peer assessment for both oral and written projects showed a similar tendency with those of the teachers in the three classes; however, follow up interviews revealed students' uncomfortable feelings in assessing their peers' language proficiency. The researchers also mentioned that the students in their study showed less uncomfortableness in assessing their peers' oral fluency such as pronunciation and pacing; these facets of a presentation do not greatly rely on students' language proficiency. They speculated that students would conflate other elements such as layout and organization with linguistic elements in assessing their peers' language proficiency. This is an informative point to choose the types of rating scales for a valid assessment in a course's final presentation.

Saito (2008) also reported the effect of rater training for assessing oral presentations involving both participant students and teachers in Japan. His study showed certain associations of the assessments between participant students and the teachers even in the group in which the students did not have rater training. Saito argues that rater training could play a role in reducing the number of misfitting raters including the teachers themselves; however, instructions on quality presentations alone benefit students in making a consistent rating. He also mentions the reality of classrooms where it is difficult to take a considerable amount of time for rater training.

Research on peer assessment has also been reported from universities in Iran over the past 10 years. Ahangari, Rassekh-Alqol, and Hamed (2013) replicated Patri's study using the same assessment criteria for assessing students' oral presentations in a university in Iran. They also found that there was a high association (r=.87) between peer and teacher assessments in the experiment group in which a training session was implemented. Furthermore, their research indicated that students gave more lenient points to their peers than their teacher; however, the difference was not statistically significant.

In a very recent study conducted in Iran, Nejad and Mahfoodh (2019) examined the associations between self-, peer, and teacher assessments of oral presentations targeting 60 Iranian students and four teachers. There were four presentation sessions with 15 students

each. The students and the teacher assessed students' 10-minute-long oral presentations using an analytic scoring rubric. Both the students and the teachers attended the rater training sessions and understood the criterion and scaling of the rubrics beforehand. As a result, a statistically significant correlation between the peer and the teachers' assessments was found. Their study also revealed the there was a statistically significant difference between the two assessments; the teachers' assessments were significantly lower than those given by the students.

Turning to the Japanese context, studies regarding peer assessment in English classrooms examined the consistency of students' rating behaviors with analyses of the many-faceted Rasch measurement (MFRM). Tanaka (2017) explored how students' personality traits affect their rating behaviors for assessing their peers' oral presentations. Her study revealed that some personality traits such as dependency on others became a predicting factor for peer rating behaviors of oral presentations as time went by. The findings of this study should be recognized especially when incorporating peer assessment into summative assessment.

A study conducted by Matsuno (2017) used peer assessment as a summative assessment in an oral presentation course targeting engineering major students. After having the students understand the key domains of analytical scaling rubrics over three successive classes as rater training, students, who became both presenters and raters, assessed their peers' 3-minute oral presentations. Matsuno consequently found that there were no misfitting students in the study except for one. Furthermore, her research reported that a majority of raters including the teacher herself rated the students' presentations leniently. Matsuno states that the reason for this is that her students completed their final presentations after following the teachers' explanation for quality presentations. Matsuno appraised peer assessment as having a good potential as a formal assessment, maintaining that having only one teacher in a classroom would be problematic for assessing all student presenters.

As the above studies indicate, peer assessment has been proven to be comparable to assessment by teachers when students are well aware of the assessment criteria and rating scales after being properly trained. Additionally, as Nejad and Mahfoodh (2019) noted, the homogenous contexts in terms of students' education levels and cultural background also contribute to the similarity of the assessment by the teacher. It is, therefore, assumed that peer assessment can be incorporated into the summative assessment along with teacher assessment in EFL college English classes.

However, it is not clear yet to what degree peer assessment should be integrated with

teacher assessment for assessing final presentations. Moreover, it is worth examining how teachers and students can consistently assess the students' presentations in case multiple sessions are required to use for completing all the students' presentations. As such, there is still room for adding to the volume of literature on peer assessment studies in the EFL classrooms for elucidating the features of the assessment.

3. Purpose of the Study

This study can be classified as action research, with the stated aim of finding a solution for a possible problem in the classroom without rigorously controlling the research site (McMillan & Schumacher, 1997). The current study is designed to explore how peer assessment should be incorporated into summative assessment for final oral presentations while examining the validity and reliability of peer assessment. From the reviewed previous empirical studies in the EFL context, it is clear that rater training helps the student raters assess their peers' presentations appropriately. It is, therefore, hypothesized that peer assessment is agreeable to use as part of summative assessment in the same way as the one by the teacher in a series of final presentations. To examine this hypothesis, the following four research questions are posted to guide the study:

R1. How consistently can students and the teacher assess their peers' /her students' oral presentations?

R2. Are there any statistically significant differences in the assessments by the students and the teacher in a series of individual presentations?

R3. Are there any statistical differences between the peer and the teacher assessments?

R4. To what degree of the association is there between the peer and the teacher assessments?

The first research question examines how reliable peer and teacher assessments are respectively. It is important to find out to what extent peer and teacher assessment are reliable for incorporating them into summative assessment. The remaining three questions (R2~R4) are designed to explore how valid peer assessment is in a comparison with teacher assessment. It is expected to be analyzed to what degree students can accurately assess their peers' oral presentations using their teacher's assessment as a benchmark if any.

4. Method

This study was conducted in a large national university in the western part in Japan in 2019. The university has a four-term system with an eight-week long term. Classes are held

once a week for 120 minutes, lasting 16 sessions. The course in this study was conducted in one of the required English classes for second-year students and designed to acquire academic listening and speaking skills. The participant students took a four-skill GTEC Academic test in December of the previous academic year in order to stream the classes in the second year's required classes. The mean scores of the GTEC of the target class was 495, which is equivalent to A2 in the CEFER (Benesse i-Career, 2018). The class consisted of 31 students whose majors engineering, chemistry, agriculture, and environmental studies.

A textbook, *Lecture Reading 1* (Oxford), was used throughout the course, and group presentations consisted of about four students were conducted at the end of Term 3. Their group presentations were assessed by the teacher who is the author of this paper referring to a 5-point rubric. The rubric was developed by the English program at the university including the following key elements of oral presentations: Content, Structure/Organization, Language, Delivery, and Visual aids. These five components were introduced as speaking strategies in the textbook.

For the final project of the course, all registered students had to conduct an eight to 10-minute oral presentation individually in front of their peers and the teacher using visual slides such as Office PowerPoint. The current study was conducted using the students' final presentations which were carried out over the last three sessions of the course with about 10 students' presentations per session. Two students were deleted from the participants of the study since they missed one of the three sessions. Consequently, the participants of this study were 29 students who attended all three presentation sessions and took on the roles of both speakers and raters.

The students chose the topics of their presentations themselves based on their interests, which included topics such as the safety of autonomous driving and the new functions of smart phones. Because of the aims of the course, the students were required to include three references in order to make their presentations academic. Prior to the final presentation sessions, a model presentation regarding home economics in a women's university was performed by the teacher in front of the students. Her presentation was used as a model for students to follow, and was further used for a rater training activity. The reason why the teacher's presentation was used as a model was both for education and practicality. Monitoring their teacher's presentation, the students could review what they had learned throughout the course. Furthermore, her presentation could be used for checking the classroom equipment such as personal computers and projector in order to make the students' presentations smooth.

The students were asked to rate the teacher's performance using the in-house rubric, which was the same as the one used for the group presentations in Term 3. The rubric used this time was, however, modified to a 10-point scale describing the quality of presentations as follows: 10 (excellent), 9-8 (very good), 7-6 (fair), 5-3 (weak), and 0-2 (unacceptable). As Berger (2015) explains, rating scales function as a common yardstick for raters to judge learners' performances, thus ensuing the reliability and validity of scores. Therefore, clear descriptors are necessary for making the student raters' judgement explicit. The teacher explained each criterion of the rubric to the students carefully before and after her presentation.

The students were also informed of the grading system awarded for the presentations in the training session. They were told that the mean scores given by their peers would account for 10 % of the total scores of the individual presentation; in other words, 90 % of the assessment was decided by the teacher. The author referred to Saito (2008) when deciding what percentage of the contribution of the peer assessment to use for this study. In Saito's (2008) research, 15% of the total scores of the individual presentations was allotted to the peer assessment. As the current study did not implement as rigorous a rater training program as utilized in Saito's study, the amount of the peer assessment that contributed to the total scores was reduced to 10%.

Students were instructed to give a single score to individual peer presenters right after his or her presentation finished. Student raters inputted the score using a *questionnaire* function on the university Moodle system with their mobile phones. The rubric used in the study can be categorized as holistic scaling. Compared to analytical scoring, holistic scoring is considered more appropriate and practical for summative and for a large number of evaluations (Weigle, 2002). Although this was a small-scale classroom study, the teacher in this study wanted her students to concentrate on their peer's presentations as audience members rather than focusing too much on calculating the scores for each criterion during the presentations. Again, this study is action research which was conducted for improving classroom practice rather than building an assessment theory.

After all the students finished their presentations, the teacher downloaded each of the students' scores given by their peers and calculated the mean scores. The final score which was converted to 100 points, including the teacher's assessment, was informed to the students individually through the Moodle system.

After organizing the scores given by the students and the teacher, statistical analyses were conducted using the FACET computer program 3.80 and IBM SPSS Statistics 22.0.

5. Results and Discussion

The results will be reported in the order of the research questions.

R1. How consistently can students and the teacher assess their peers' /her students' oral presentations?

Consistency was analyzed using the FACET computer program. MFRM can provide information on the rating behavior of each rater based on the different factors, called facets. In this study, there are three facets: student speakers' ability, holistic scoring rubric, and raters' behavior.

Before the analyses of MFRM, the original 10 rating scales were converted into 5 ratings. According to Linacre (1999), each category needs at least 10 ratings; otherwise, the findings may be unstable. In the original 10 rating scales, there were no counts used for scores of 1, 2 and 3. Further, a score of 4 and score of 5 had 2 and 6 counts respectively, violating Linacre's suggestion. Therefore, the scales from 1 to 6 were combined together, resulting in a 5-point scale, as shown in Table 1. In the 5-rating scale, as the scales are increased, the logits also become larger. Additionally, outfit mean-squares of each converted scale are all near 1.0, which meets Linacre's criteria, showing that observations fit with their categories. Therefore, further MFRM analyses were undertaken with the 5-point scale.

Table 1. Category Statistics

Original score	Converted scale	Counts used	%	Logit (Average measurement)	Outfit MnSq
1+2+3+4+5+6	1	63	8	-2.63	.8
7	2	151	18	-1.40	.9
8	3	297	36	.06	1.0
9	4	214	26	1.64	1.1
10	5	110	13	3.15	1.2

Figure 1. Variable map of MFRM analysis. Each asterisk (*) indicates one student participant. T in *Raters* indicates the teacher in the study.

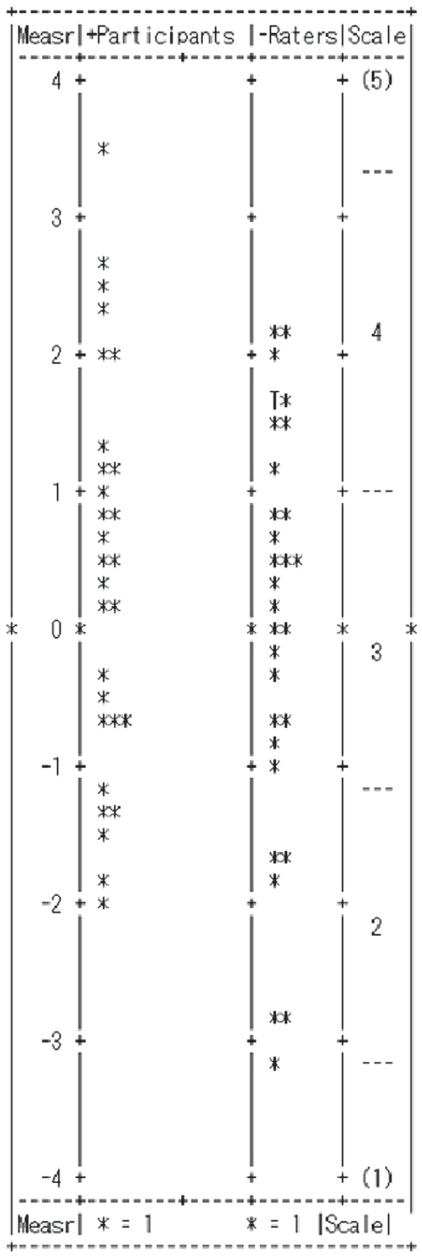

Figure 1 above reports the relationships among participant speakers' ability, rater severity, and scales. The first column indicates the Rasch logit scale. The logit scale shows the same measurement scale unlike the same distances between scores in the observable rating (Eckes, 2015).

The second column in Figure 1 shows the participant speakers in this study. Each asterisk represents one speaker. The participants are distributed on the map according to their presentation abilities with the highest at the top and the lowest at the bottom. As shown in Table 2 below, the logits varied from -1.92 to 3.55. The mean squares of infit and outfit are .98 and 1.02 respectively. According to Linacre (2012), mean squire values of productive for measurements are from .5 to 1.50. The ability of the participants' speaking ability meets Linacre's criteria showing an acceptable range.

Table 2. Descriptive Statistics of Participants and Raters

Facets	N	M(SD)	Min to Max	Range	Separation	Infit MnSq (ZStd)	Outfit MnSq (ZStd)	Strata	Reliability
Participants	29	3.19(.62)	-1.92 to 3.55	5.47	4.86	.98 (-.4)	1.02 (.4)	6.81	.96
Raters	30	3.19(.64)	-3.18 to 2.12	5.3	4.99	.99 (-.2)	1.02 (-.1)	6.99	.96

The participation separation value was 4.86, meaning that the speakers in this study can be separated into about five levels of speaking abilities. The participants' reliability showed .96, which is the Rasch reliability equivalent to the Cronbach Alpha statistics.

The third column in Figure 1 shows rater severity. The teacher, who is the author of this paper, is also included in the MFRM analysis. *T* indicates the teacher and is noticed as the fourth severest rater (1.72 logits) among the 30 raters. The severest rater (2.12 logits) is at the top in the column, and the most lenient rater (-3.18 logits) is at the bottom. The rater separation value was 4.99, which shows about five levels of rating severity with a reliability of .96. As also shown in Table 2, the mean squire values of infit and outfit are .99 and 1.02 respectively, meaning that the raters' fit statistics met Linacre's criteria, too.

From the measurement reports of Figure 1 and Table 2, it is noted that raters in this study gave scores to their peers and her students somewhat leniently for the participant speakers, especially for the lower peer speakers, while the ranges of logits of participants and rates are similar with 5.47 and 5.3 respectively. This result echoed previous studies (e.g., Matsuno, 2017; Nakamura, 2002; Patri, 2002; and Saito, 2008).

Table 3. Percentage of Mean-Square Fit Statistics

Facets	N		Overfit (value<.60)	Fit (.50≤Value≤1.5)	Underfit (1.5≤Value)	Distorts or degrades the measurement system. (Value>2.0)
Participants	29	Infit	0.00	100	0.00	0.00
		Outfit	0.00	93.10%(N=27)	6.89%(N=2)	0.00
Raters	29	Infit	17.24%(N=5)	68.97%(N=20)	13.79%(N=4)	0.00
		Outfit	17.24%(N=5)	65.52%(N=19)	17.24%(N=5)	3.44%(N=1)
Teacher	1	Infit		100 (.71 -1.2)		0.00
		Outfit		100 (.70 -1.2)		0.00

Table 3 shows the results of mean-square fit statistics by MFRM for the participant speakers and raters. As indicated in the table above, the teacher's assessment was noticed as consistent, meeting the criteria both for infit and outfit statistics: .71 to 1.2, and .70 to 1.2 respectively. On the other hand, approximately 70% of the student raters in this study met the criteria for infit statistics and 65% for outfit statistics. While the infit statistics indicate the rater's assessment pattern or consistency, the outfit statistics are affected by outlying factors such as the speakers' abilities and the evaluation criteria. It is commonly recognized that infit is more important to judge the reliability of raters than outfit (Eckes, 2015). One of the aims of this study is to examine how raters can consistently provide their peers and the students with the scores; therefore, infit statistics are prioritized for analyses as an indicator for fit statistics. Consequently, it is interpreted that 70% of the participant students can consistently assess their peers' oral presentations.

Moreover, according to Linacre (2012), the values beyond 2.0 have the possibility of distorting or degrading the measurement model so that those raters should be eliminated from the assessment. In this study, however, there were no students whose rating logits exceeded the value of 2.0. In other words, all participant students in this study can contribute to the productive mean scores for assessing their peers' final oral presentations.

Research questions from *R2* to *R4* were analyzed by IBM SPSS Statistics 22.0. The results of the analyses will be discussed in the order of the questions.

R2. Are there any statistically significant differences in the assessments by the students and the teacher in a series of individual presentations?

Table 4 below reports on the descriptive statistics of peer and teacher assessments in the

three presentation sessions.

Table 4. Descriptive Statistics of Students' and Teacher's Assessments in the Three Sessions

								95% Confidence Interval							
								Lower		Upper					
Sessions	N	Mean		S.D.		S.E.		Bound		Bound		Minimum		Maximum	
		S	T	S	T	S	T	S	T	S	T	S	T	S	T
1	9	8.08	7.11	.68	1.69	.22	.56	7.55	5.81	8.61	8.41	7.08	4.00	9.22	9.00
2	11	8.21	7.18	.60	1.16	.18	.35	7.80	6.39	8.61	7.96	7.31	5.00	9.07	9.00
3	9	8.33	7.55	.73	.88	.24	.29	7.77	6.87	8.89	8.23	7.13	6.00	9.55	9.00
Total	29	8.21	7.27	.65	1.25	.12	.23	7.96	6.80	8.45	7.75	7.08	4.00	9.55	9.00

Note. S indicates the participant students in this study; T indicates the teacher in this study.

As shown in Table 4 above, the mean scores given by the students are higher than those given by the teacher in all three sessions. It is also noticed that the scores increase as the session progresses for both groups.

A one-way ANOVA was conducted to measure the mean differences of the students' assessments in three sessions after the normality test. As a result, there were no statistically significant differences between the mean scores in the three sessions (F =.311. df = 2, p = .736).

Regarding the teacher's assessment, a Kruskal-Wallis test was utilized since it did not follow a normal distribution. The test did not show statistically significant differences between the mean scores in the three sessions, either. Given the results of those two tests, both peer and teacher assessments did not show statistically significant differences between the three sessions.

R3. Are there any statistical differences between the peer and the teacher assessments?

Table 5 presents the comparisons between the students and the teacher assessments. To examine the differences between the two assessments, a Wilcoxon sighed-ranks test was conducted. As indicated in Table 5 below, there was a statistically significant difference between the two assessments; z = −4.563, p = .000, and r = .85. This means that the teacher's assessment is significantly severer than the assessment by the students in this study.

Table 5. Comparison Between Peer and Teacher Assessments

	N	Mean	S.D.	Min.	Max.	*Z*	*p*	*r*
Teacher	1	7.27	1.25	4.00	9.00	−4.563	.000	.85
Students	29	8.21	.65	7.08	9.55			

The degree of severity between the peer and teacher assessments varies from study to study. A study conducted by Nejad and Mahfoodh (2019) shows a similar result to this study; the teachers' group's mean scores were lower than those by the students. Ahangari, Rassekh-Alqol, and Hamed (2013) also found that the teacher's assessment was slightly severer than the peer assessment even though there were no statistically significant differences between them. On the other hand, a study conducted in a university in Japan by Nakamura (2002) indicated that teacher assessment was more lenient than the assessment by students. Matsuno's study (2017) also showed a lenient tendency for the teachers.

As shown from the review of the handful of studies above, it should be mentioned that these studies used an analytical scoring rubric which sums up to a total score to award the students' presentations. Some researchers (e.g., Hamp-Lyon, 1991) see a high reliability with analytical scoring rubrics for raters to limit to a single construct. This study, on the other hand, used a holistic type of scoring rubric to give a single score. As reported in Research Question 1, the students in this study awarded the scores to their peers in a consistent manner. It can be concluded, therefore, that the rubric in this study seems to have functioned well. The students in this study were expected to take the role of judge more than as a member of the audience. Despite this initial expectation, there were statistically significant differences in rating severity between the students and the teacher. Considering those points, although the students and teacher in this study had a common understanding of the assessment criteria, their expectation toward the goal of the final presentations might be very different from each other.

R4. To what degree of the association is there between the peer and the teacher assessments?

Table 6 indicates the results of correlation analysis between the mean scores given to individual students by the student raters and the scores to individual students by the teacher.

Table 6. Correlation Between Peer and Teacher Assessments

	Teacher (N=1)
Peer (N=29)	.676**

Note. ** indicates statistically significant at .0001 level(two-tailed).

Kendall's tau-b(τ) correlation was chosen to measure the strengths of associations between the two assessments considering the small population size and non-parametric features of the teacher's assessment. As seen in the table above, Kendall's tau-b value shows .676, which is statistically significant at .0001. This value is almost the same as the one which was reported by a meta-analysis conducted by Saito (2008). He reviewed 56 studies and calculated the mean scores of the Pearson's correlation coefficient as r = 69 (r^2 = .47). Recalling the more recent studies on peer assessment introduced earlier, the value of this study was found to be the lowest. For example, a similar classroom study by Tanaka (2017) showed a Pearson's correlation coefficient value of .82 (p <.001). The current study, however, shows a .80 (p <.001) Pearson's correlation coefficient. Since the teacher's assessment did not follow a normal distribution as mentioned above, Kendall's tau-b(τ) correlation was chosen for correlation analysis. Considering the values of the strengths of association between the two assessments, this study also indicated a strong value. Unlike in those studies, however, peer training was incorporated in the regular session right before the three sessions of students' presentations. Despite this, the association of the two assessments is found to be considerably strong.

6. Conclusion and implications

This action study echoed previous studies on comparisons between peer and teacher assessments in the EFL contexts. Implementing rater training in the regular classroom session was proven to help the students rate their peers' oral presentation consistently in a valid manner.

Although it is not surprising that there is considerable agreement between peer and teacher assessment, this study indicated the differences of severity between the two assessments. Patri (2002) mentions that peer and teacher assessments still have different features despite a strong association with each other even with firmly set rating scales. Additionally, Cheng and Warren (2005) investigated students' attitudes towards peer assessment and found that students might not consider all the elements of assessing criteria when awarding scores to their peers' presentations. These points were reflected in the current study.

This study tasked raters with awarding a single score to their peers while referring to a holistic rubric with five domains. Students should have judged the whole quality of their peers' presentations without paying attention to a single criterion. It is, on the other hand, speculated that teachers assess presentations more objectively than students when considering those domains to make sure to what extent his or her teaching goals were achieved in the final

presentations. Thus, peer and teacher assessments still have different characteristics. Perhaps, while students are still emphasized in the role of audience members, the teacher can show a more explicit role as a judge. However, both roles are still important when considering the dynamics of classroom culture.

The question still remains as to the use of peer assessment for summative assessments: to what extent is it incorporated into the final scores? Considering the results of this study, it might be suggested to increase it to more than 10% of peer assessment when incorporating it into the final grade. Of course, however, it should be considered from the view point of the friendship effect. Nakamura (2002) points out that not only students, but teachers tend to avoid giving lower scores to the presenting students. If students are informed of a large ratio of their assessment, such as a 50-50 rating between teachers and students, the friendship effect will be intensified.

Additionally, according to Tanaka (2017), the personal traits of the students might also affect their rating behavior. It is, therefore, very difficult to present an exact ratio, but it is safer to say that it should not be beyond 30% of the total score considering that 30% of the students were not within the range of infit statistics in this study. A majority of students in this study might have felt relaxed to present and could seriously score their classmates' performances since they were told that their contribution to the total assessment would only account for 10 % of the total score. If they were told that their contribution to the assessment would account for a half of the total score, they might have felt pressure, and in return give higher scores to their peers because of their friendship with them.

Additionally, in this study, a teacher's model presentation was used for the purposes of rater training. As far as the studies reviewed for this paper, there were no studies to be found where the teacher's model presentation was used for rater training; most of them used videos of previous students' presentations. From a sense of practitioners, teachers usually show their model writing or speech when assigning tasks to students. On the other hand, it would be motivating for students to see models by previous students for rater training. However, a consent form is required in advance to protect students' personal information when videotapes of previous students' presentations are used. Also, presentations from previous years might not perfectly fit the presentation contexts for the target students. As a result, teachers' models are considered more practical and effective in the classroom.

The teacher, the author of this paper, in this study was a non-English speaker teacher of English. Showing her presentation in English was also challenging to her. Through her model

presentation, however, she was hoping that students would be encouraged and become confident in their presentations, realizing that there will be no extreme point from zero mastery to full proficiency, but a specific range between the two extreme poles (Berger, 2015).

This study was conducted in a moderate size general English classroom in an EFL context. Rater training was incorporated into regular classes as a classroom activity. The results of this study might not be applied to other EFL contexts. However, this study will add to the volume of literature of peer assessment showing that it has a good potential to become part of summative assessment. A single teacher's assessment is problematic for assessing students' output skills in the classroom.

References

Ahangari, S., Rassekh-Alqol, B., & Hamed, L. A.A. (2013). The effect of peer assessment on oral presentation in an EFL context. *International Journal of Applied Linguistics & English Literature, 2* (3), 45-53. https://doi.org/10.7575/aiac.ijalel.v.2n.3p.45

Benesse, i-Career, (2018). 学生の4技能英語力育成に向けた大学英語教育の在り方とは【What is the role of university English education in fostering students' proficiency in the four skills of English?】. Benesse

Berger, A. (2015). *Validating Analytic Rating Scales*. Frankfurt: Peter Lang.

Cheng, W. and Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22* (1), 93-121. https://doi.org/10.1191/0265532205lt298oa

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement.* Frankfurt: Peter Lang.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 241-276). Norwood:Ablex.

Linacre, J. M. (1999). Investigating rating scale category utility. Journal of Outcome Measurement, *3* (2), 103-122. Retrieved from http://jampress.org/jom_v3n2.pdf

Linacre, J. M. (2012). Many-Facet Rasch Measurement: Facets Tutorial. Retrieved from http://winsteps.com/tutorials.htm

Matsuno, S. (2017). Adoptability of Peer Assessment in ESL Classroom. *Creative Education, 8*, 1292-1301. https://doi.org/10.4236/ce.2017.88091

McMillan, J. H., & Schumacher, S. (1997). *Research in education*. New York: Longman.

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

Nejad, A. M., & Mahfoodh, O. H. A. (2019). Assessment of oral presentations: effectiveness of self-, peer-, and teacher assessments. *International Journal of Instruction, 12* (3), 615-632.

https://doi.org/10.29333/iji.2019.12337a

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies 44*. 203-215.

Orsmond, P., & Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21* (3), 239-250. https://doi/abs/10.1080/0260293960210304

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing, 19* (2) 109-131. https://doi.org/10.1191/0265532202lt224oa

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25* (4) 553–581. https: //doi.org/10.1177/0265532208094276

Tanaka, M. (2017). Examining personality bias in peer assessment of EFL oral presentations: A preliminary study. *JALT Journal, 39* (2). 183-196. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj2017b-research.pdf

Villamil, O.S., & de Guerrero, M. C. M. (2006). Sociocultural theory: A framework for understanding the social-cognitive dimensions of peer feedback. In K. Hyland & F.

Hyland (Eds.), *Feedback in Second Language Writing: Contexts and issues* (pp.246-265). Cambridge: Cambridge University Press.

Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press.