The Japan Language Testing Association

Applicability of Peer Assessment for Classroom Oral Performance

Akiyo HIRAI
University of Tsukuba
Naoko ITO
Graduate School, University of Tsukuba
Toshihide O'KI
Hakuoh University

Abstract

Peer assessment is a practical tool for in-class performance testing. In order to investigate its applicability for oral performances, this study examines the extent to which peer assessments by Japanese students correlates with teacher assessments in terms of the following conditions: (a) anonymity of the rater and (b) peer discussion of students' oral performances prior to assessment. Students' oral performances in story retelling tasks were rated by their peers and teacher and these assessments were subsequently compared. The results showed that (1) peer and teacher assessments were significantly correlated when anonymity of raters was preserved and (2) determining ratings in pairs did not increase the correlation between peer and teacher assessments. These results suggest that whether peer assessment is a genuine tool for evaluation remains open to question.

Key Words: peer assessment, oral performance, prior discussion, anonymity

Introduction

With the growing interest of English teachers in Japan in fostering students' English communication abilities, particular attention has been directed toward speaking ability as an important component of communication. However, teachers are often reluctant to assess students' speaking abilities because doing so is time consuming. For example, in order to evaluate all students in a class of 40, a teacher has to interview each individual in a direct face-to-face oral situation, or listen to performances recorded by students individually during an indirect, tape-mediated oral test administered in the classroom. In order to alleviate their burden and use class time more effectively, teachers may consider using the method of "peer assessment" or "peer evaluation" by students.

Saito (2008) describes the characteristics and benefits of peer assessment in the

NII-Electronic Library Service

The Japan Language Testing Association

following manner: (1) students' ratings have been found to strongly correlate with those of the teacher, (2) peer assessment can help students reflect on their own learning, (3) students often respond to peer assessment positively, and (4) assessment can help students develop a sense of "shared responsibility" (p. 554). Although these advantages of peer assessment are conceivable, they remain inconclusive and further investigation is needed, particularly with regard to item (1) above. Thus, the purpose of this study is to investigate whether peer assessment can be sufficiently reliable to replace teacher assessment. As compared to other tests, such as entrance exams or term tests, peer assessment is not usually considered a high-stakes test. However, depending on the degree to which peer assessment scores will be used for grading purposes, it may affect students' final grades. Thus, teachers must understand the nature and substitutability of peer assessment for teacher assessment.

Issues of Peer Assessment

With regard to the applicability of peer ratings, a number of studies seem to indicate that a teacher's rating of oral tests can be replaced by a student's rating because a high correlation between peer and teacher ratings indicates that peer assessment can be as reliable as teacher assessment (e.g., AlFallay, 2004; Campbell, Mothersbaugh, Brammer, & Taylor, 2001; Fukazawa, 2009; Hughes & Large, 1993; Langan et al., 2008; Miller & Ng, 1994). However, there are also studies that have failed to find a strong correlation between the two types of rating (e.g., Freeman, 1995; Jafapur, 1991).

These conflicting results might reflect the differences in assessment conditions and students characteristics among these studies. Table 1 summarizes results of nine studies assessing students' oral performance (e.g., oral tests and presentations). The results differ from each other in terms of (a) the language on which the study focused (L1 or L2), (b) whether peers participated in discussion prior to assessment, (c) whether the final rating was obtained by averaging multiple scores (single rating or mean rating), and (d) the number of students involved (20 to 210).

With regard to L1/L2 differences, peer assessment in L2 settings is assumed to be less reliable than in L1 settings because it is more challenging for L2 learners to provide accurate evaluations. The lowest two correlations (i.e., r = .44 and .49) were both from L2-focused studies. When assessing L2 oral performance, the rater often has to process even unfamiliar linguistic aspects of the performance in order to comprehend the content. As a result, L2 learners, who have much less linguistic knowledge in the target language than native speakers, could make inconsistent judgments regarding their peers' performance. Hence, this study considered it meaningful to focus on peer assessments performed by Japanese learners of English.

NII-Electronic Library Service

The Japan Language Testing Association

Table 1

Summary of the Correlation Between Peer and Teacher Assessments in Nine Studies

Study	Language (Country/ Region)[a]	Prior Discussion	Single(S)/ Mean(M) Rating	Other Conditions	*N*	Correlation (*r*)
Jafapur (1991)	L2 (Iran)	No	S	Anonymous	41	.44*
Hughes & Large (1993)	L1 (UK)	Yes	M		44	.83**
Miller & Ng (1994)	L2 (Hong Kong)	Yes	S	Class 1	20	.68[b]
Miller & Ng (1994)	L2 (Hong Kong)	Yes	S	Class 2	21	.80[b]
Freeman (1995)	L1 (Australia)	Yes	M		41[c]	.60[b]
Campbell et al. (2001)	L1 (US)	No	M		60	.58*
Patri (2002)	L2 (Hong Kong)	No	S		29	.49**
Patri (2002)	L2 (Hong Kong)	Yes	S		25	.85**
AlFallay (2004)	L2 (Saudi Arabia)	Yes	S		200	.82**[d]
Langan et al. (2008)	L1 (UK)	No	M		60	.77**
Fukazawa (2009)	L2 (Japan)	No	M	Class A	36	.92**
Fukazawa (2009)	L2 (Japan)	No	M	Class B	35	.93**
Fukazawa (2009)	L2 (Japan)	No	M	Class C	35	.90**
Fukazawa (2009)	L2 (Japan)	No	M	Class D	39	.79**

Note. [a]The target language in all the studies was English. [b]The provability of the correlation was not reported. [c]Forty-one groups comprising 210 peers each rated oral presentations. [d]The values were Spearman's correlation coefficients for ranked data (r_s). *$p < .05$, **$p < .01$.

Despite the above findings, some L2-focused studies have shown rather high teacher-peer correlation (e.g., AlFallay, 2004; Patri, 2002). This may be due to the effect of engaging in discussion prior to assessment. That is, in the studies listed in Table 1, there seems to be a tendency toward higher correlations when students rated a peer's performance after discussion (e.g., r = .83, .80, .85, and .82) than when students assigned a rating without engaging in prior discussion (e.g., r = .44, .58, and .49). As indicated by Patri's (2002) study, even L2 students can show a high peer-teacher correlation (r = .85) when they engage in group discussion prior to assessment. This may be because students' subjectivity is minimized through discussion with peers. However, in Fukazawa's (2009) study, peers did not discuss before assigning ratings but their assessments were still highly correlated with those of the teacher's. This may have been because the rating of each student's performance was obtained by averaging all the scores of the participants, as indicated by Mean Rating (M) in Table 1. This process might be helpful for minimizing the deviation of peer ratings and increasing the correlation between peer and teacher assessments. Thus, overall, the effectiveness of prior discussion appears to vary, and it is uncertain if "discussion prior to

NII-Electronic Library Service

The Japan Language Testing Association

assessment" is an important condition for reliable peer assessment.

The next issue that seems to be relevant is "anonymity." In anonymous conditions, examinees do not know who the rater is. This condition may have both positive and negative effects. A study by Orsmond, Merry, and Reiling (1996), which is not included in Table 1 since it was not related to oral presentations but poster presentations, preserved anonymity by having the two groups of peer raters exchange rooms before they assigned ratings. The correlation between peer and teacher ratings was high, at .73, supposedly because anonymity helped to reduce the anxiety that raters felt about being accused of excess severity by their peers. In contrast, the study by Jafapur (1991), in which peer anonymity seemed to be preserved simply by not revealing the rater's name to the presenter, showed a relatively low correlation (r = .44). The difference between the results of these two studies could be attributed to the difference in task types (i.e., poster presentation versus oral interview). Since results regarding anonymity are inconclusive, further investigation using the same task types is needed in order to determine whether anonymity is indeed an important condition for reliable assessment of Japanese EFL learners.

In order to address issues of inconclusiveness, this study aimed to answer the following research questions (RQs):

RQ1. To what degree does peer assessment correlate with teacher assessment in oral testing of Japanese students of English?
RQ2. Do the assessment scores of peers differ significantly from that of the teacher?
RQ3. When the anonymity of the rater is preserved, will the correlation increase compared to when the identity is known?
RQ4. Which will show a higher peer-teacher correlation, peer assessment with prior discussion or peer assessment without it?

These RQs were investigated through two experiments. RQ1 and RQ2 were explored in both Experiments 1 and 2, whereas RQ3 and RQ4 were examined by comparing the results of Experiments 1 and 2.

Experiment 1

Method

Participants

The participants were 80 Japanese university freshmen majoring in the natural sciences, aged 18 or 19. They belonged to two classes taught by one of the researchers: Class A (n = 40) and Class B (n = 40). The experiment was conducted in two adjacent

NII-Electronic Library Service

The Japan Language Testing Association

CALL classrooms with a CALL equipment room in between. Therefore, one teacher (i.e., one of the researchers) was able to give instructions to the two classes simultaneously. Data availability was affected by absences from both tests and by recording malfunctions; however, ultimately, data from 50 students were available for analysis.

Materials

The oral test. A Story Retelling Speaking Test (SRST) developed by Hirai and Koizumi (2009) was used as an oral test. This is a tape-mediated test that requires test takers to read a story, answer questions about it orally, and then retell it prompted only by keywords from the story. The three texts for the SRST were adapted from the interview and reading sections of the pre-second grade EIKEN tests from 2005 and 2008. The Flesch-Kincaid Grade Levels of stories A, B, and C were 7.4, 6.2, and 6.1, respectively. The difficulty of the texts was considered appropriate for the participants, since almost half of them had passed either the pre-second or second grade EIKEN tests.

The EBB scoring scale. An Empirically-derived, Binary-choice, Boundary-definition (EBB) scale was prepared for scoring the SRST (see Appendix). The EBB was originally developed by Upshur and Turner (1995) and then further developed for speaking tests by Hirai and Koizumi (2008). According to Hirai and Koizumi, the latter scale was slightly superior to analytic scales in terms of validity and reliability. The scale is based on three criteria: (1) "Communicative Efficiency," which mainly measures production amount, fluency, and coherency; (2) "Grammar & Vocabulary," which focuses on grammatical accuracy and appropriate use of vocabulary; and (3) "Pronunciation," which focuses on segmental and supra-segmental features, such as pronunciation, stress, and intonation. Participants were given a rating of one to five points for each criterion.

Procedure

The study comprised two sessions: the oral test (the SRST) and the peer assessment. In the SRST session, the participants read a story while being guided by a few comprehension questions. They were then asked to retell the story and express their opinions about it, looking at keywords printed on a handout. Students' performances were recorded on tape simultaneously in the language lab. In the latter session, the students exchanged tapes with the student next to them and rated the recordings using the EBB scale. Prior to rating a peer's performances, they were given a detailed explanation of the three EBB scale criteria.

The recordings were also subsequently assessed by the researchers, and these assessments qualified as the teachers' assessments. Half the students' performances (n = 25) were rated first by two researchers in order to check the inter-rater reliability.

NII-Electronic Library Service

The Japan Language Testing Association

The correlations between the two raters for Communicative Efficiency, Grammar & Vocabulary, and Pronunciation were .84, .79, and .69, respectively. Although the last coefficient was relatively lower than the first two, we regarded it as sufficient given our small sample size; the rest of the speech performances were rated by one of the researchers.

Results and Discussion

Table 2 presents the descriptive statistics and the correlation coefficients for peer and teacher assessments, which are also illustrated in Figure 1. Overall, the mean score for peer assessment (M = 10.64) was higher than for teacher assessment (M = 8.24). The peers' mean score for Pronunciation (M = 3.67) was the highest among all the scores, while the teacher's mean for Grammar & Vocabulary (M = 2.55) was the lowest.

Table 2

Descriptive Statistics for Peer and Teacher Assessments (N = 50) and the Correlation Coefficients of the Two Assessments

	Peer		Teacher		*R*
Criteria	*M*	*SD*	*M*	*SD*	
Communicative Efficiency	3.45	0.70	2.85	0.80	.28
Grammar & Vocabulary	3.52	0.72	2.55	0.61	.19
Pronunciation	3.67	0.73	2.84	0.72	.21
Total	10.64	1.63	8.24	1.53	.17

Note. None of the correlation coefficients were significant.

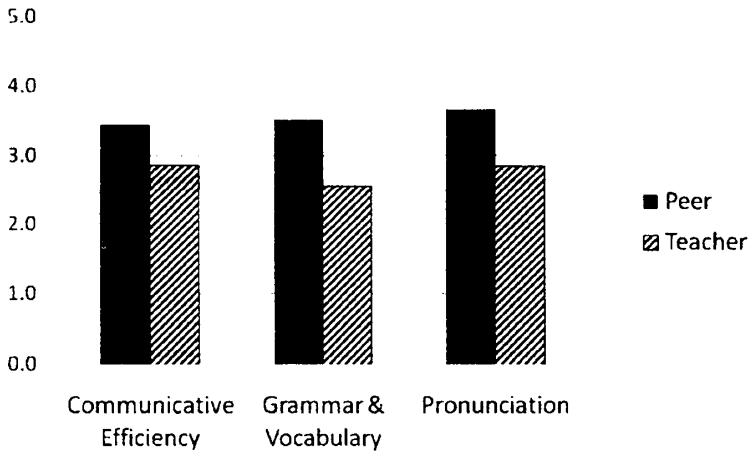


Figure 1. Experiment 1: Mean ratings of peer and teacher assessment.

As Table 2 indicates, non-significant weak correlations were observed between

NII-Electronic Library Service

The Japan Language Testing Association

peer and teacher assessments. Only the correlation regarding the Communicative Efficiency criterion was marginally significant (r = .28, p = .051), thereby implying that it was relatively easier for students to assess Communicative Efficiency than the other two criteria. This may be because Communicative Efficiency primarily measures the fluency of students' performances, such as the frequency of pauses and the amount of retelling and opinions, and thus does not demand a high level of linguistic knowledge on the part of the rater. The other two criteria should be more difficult to assess because they demand that raters have comprehensive knowledge of English grammar, vocabulary, and pronunciation.

In order to examine if there were any significant differences between peer raters and the teacher and among the three criteria (i.e., Communicative Efficiency, Grammar & Vocabulary, and Pronunciation), a two-way repeated ANOVA measure with a 2 × 3 design (Rater [Peer, Teacher] × Criterion [Communicative Efficiency, Grammar & Vocabulary, Pronunciation]) was used. The results showed that the interaction between rater and criterion was significant, $F(2, 98) = 3.52$, $p = .033$, $\eta^2 = .01$, thereby indicating that the two raters assessed each criterion differently. In all the criteria, peers' ratings were significantly higher than those of the teacher ($p < .01$). The greatest mean difference was 0.97 in the Grammar & Vocabulary section.

Moreover, *post hoc* tests revealed that the differences among the three criteria were significant in teacher assessment; $F(2, 98) = 3.96$, $p = .022$, $\eta^2 = .07$. Multiple comparisons revealed that the teacher had rated Grammar & Vocabulary ($M = 2.55$) significantly lower than Communicative Efficiency ($M = 2.85$; $p = .020$), whereas the peers rated all the criteria almost equally. Therefore, it could be said that the teacher rated students more severely than the peers, particularly on Grammar & Vocabulary.

These results may be attributable to three possible causes. First, it was likely that peers were more generous in assessing the performances of their classmates than the teacher because they did not want to be harsh for fear of damaging their relationships with their classmates. This implies the importance of rater anonymity. Second, students are incapable of identifying errors if they lack the knowledge necessary to identify them. One method that might be used to compensate for students' lack of linguistic knowledge is to provide students with the opportunity to discuss the rating process with their peers. Third, the EBB scale might have been a new and unfamiliar rating method for students, and/or it might have been too difficult to separate performance into five levels within each criterion. In this case, rater training is necessary until raters become accustomed to using the method. In addition, simplifying the scale to include three or four levels within each criterion might allow raters to feel more confident about their ratings.

NII-Electronic Library Service

The Japan Language Testing Association

Experiment 2

The results of Experiment 1 suggest that the low correlations between peer and teacher assessments were caused by anonymity, opportunities to have prior discussions, and rater training. Therefore, these assessment conditions were included in Experiment 2.

Method

Participants

The participants were drawn from two classes like Experiment 1. This time, 60 students were included in the study. In order to determine whether two classes (Group A, $n = 26$; and Group B, $n = 33$) could be treated as equivalent, a proficiency test was conducted. This test comprised three sections with a total of 134 questions: (a) a multiple-choice grammar test (16 questions), (b) a written vocabulary test (78 questions), and (c) an oral vocabulary test (30 questions). The test included a large number of vocabulary questions in Sections b and c, since these same questions had been used in different studies and worked well for identifying participants' oral proficiency levels (e.g., Hirai & Koizumi, 2009; Koizumi, 2005; Yamashita, 2008). The participants' average vocabulary size was calculated from the vocabulary tests, which were at the 2000-word level. Since the proficiency test's scores for the two classes and for participants in Experiment 1 were not significantly different, $F(2, 106) = 0.53$, $p = .589$, $\eta^2 = .01$ (see Table 3 for their proficiency scores), the two classes were treated as equivalent groups and were compared with the participants of Experiment 1.

Materials

The oral test. An SRST was conducted using two stories that differed from those used in Experiment 1. Two stories with different lengths but similar difficulty levels, which had worked well in a pre-test, were prepared: one was a 98-word passage (Story D) from an interview test used in the 1992 third grade EIKEN test and the other was a 153-word passage (Story E) from a reading test used in the 2001 fourth grade EIKEN test. The Flesch-Kincaid Grade Levels of Stories D and E were 4.1 and 4.5, respectively.

Questionnaire. In addition, a questionnaire comprising 13 questions was prepared in order to investigate student attitudes toward participating in the SRST and peer assessment. It asked students about (1) their attitude toward peer assessment and anonymous assessment, (2) their sense of responsibility toward peer assessment, and (3) the washback effect of peer assessment on language learning. The students rated each element using a 6-point Likert scale from 1 'Strongly Disagree' to 6 'Strongly

NII-Electronic Library Service

The Japan Language Testing Association

Agree.'

Procedure

Both Groups A and B took the SRST that contained the two stories explained in the *Materials* section. Thereafter, they participated in peer assessment in anonymous conditions, which differed from the condition in Experiment 1 (see Table 3). In order to ensure the anonymity of the raters, the recorded tapes were mixed and exchanged between the two groups prior to peer assessment. As illustrated in Figure 2, the students in Group A discussed the ratings of other students with the students sitting next to them, whereas the students in Group B worked completely independently, as in Experiment 1. The instructor explained the criteria for the EBB scale in detail for approximately 15 minutes, having the students listen to recordings of some benchmark performances. Since the teacher raters in Experiment 1 were the same as in Experiment 2 and the interval between the two experiments was short, only one teacher rated all the recordings.

Table 3

Proficiency Scores and Conditions for the Three Groups in Experiments 1 and 2

Group	*n*	*M*	*SD*	Discussion	Anonymity
Experiment 1	50	79.52	7.82	No	No
Experiment 2: Group A	26	79.04	6.76	Yes	Yes
Experiment 2: Group B	33	77.36	12.93	No	Yes

Note. Group A: peer assessment after discussion in pairs; Group B: peer assessment conducted independently.

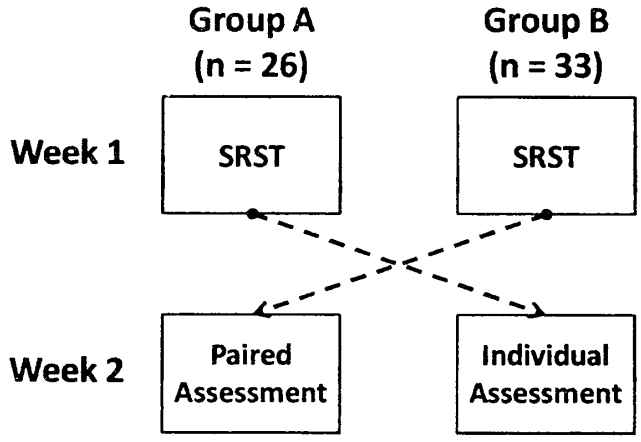


Figure 2. Procedures for peer assessment in Groups A and B. Dotted arrows indicate the exchange of the tapes on which performances were recorded.

NII-Electronic Library Service

The Japan Language Testing Association

Results and Discussion

Table 4 presents the descriptive statistics for the ratings of the two groups, and these are also illustrated in Figure 3. As was the case with Experiment 1, the peer ratings in both groups were slightly higher than those of the teacher.

Table 4

Descriptive Statistics for Peer and Teacher Assessments of Groups A and B

	Peer		Teacher	
Criteria	*M*	*SD*	*M*	*SD*
Group A (n = 26)				
Communicative Efficiency	3.08	0.80	2.93	0.76
Grammar & Vocabulary	3.12	0.79	2.67	0.66
Pronunciation	3.27	0.59	2.81	0.81
Total	9.46	1.54	8.40	1.65
Group B (n = 33)				
Communicative Efficiency	2.95	0.79	2.76	0.93
Grammar & Vocabulary	2.79	0.73	2.11	0.86
Pronunciation	3.26	0.85	2.68	1.06
Total	9.00	1.39	7.55	1.68

Note. Group A: peer assessment after discussion in pairs; Group B: peer assessment conducted independently.

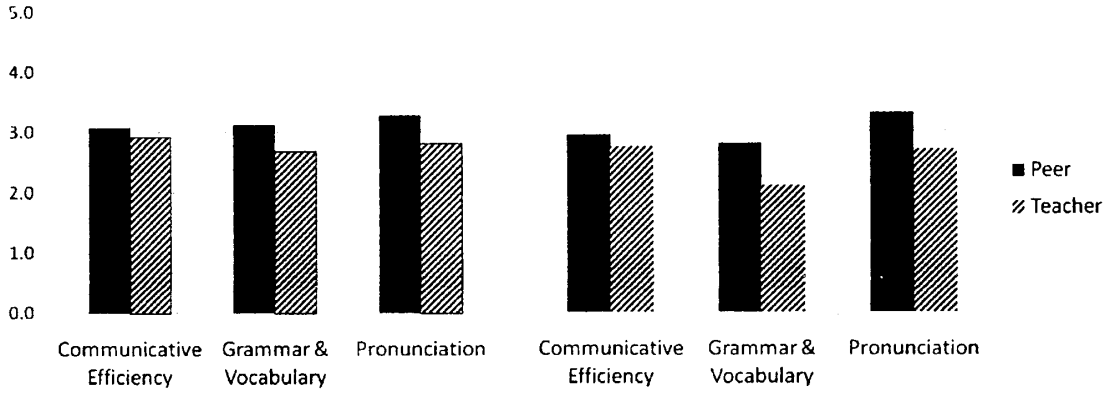


Figure 3. Average scores from peer and teacher assessments. Group A (after discussion, left); Group B (independently, right).

In order to investigate the effects of rater anonymity (RQ3), the Pearson's correlations between peer and teacher assessments in the two experiments were compared (Table 5). The results show that, even in cases where no discussion between

NII-Electronic Library Service

The Japan Language Testing Association

peers occurred (i.e., Experiment 1 and Group B), correlations were much higher when anonymity was preserved (r = .63 in Group B) than when anonymity was not preserved (r = .17 in Experiment 1). This suggests that the anonymity of raters played a significant role in increasing the reliability of peer assessment. However, interestingly, the effect of anonymity disappeared when ratings were made in pairs, as shown by the fact that Group A had the lowest correlation (r = .06). In other words, the anonymity was only effective when raters worked independently.

Table 5

Correlations between Peer and Teacher Assessments in Experiments 1 and 2

Criteria	Experiment 1 assessment: individual (n = 50)	Experiment 2 (Group A) assessment: pairs (n = 26)	Experiment 2 (Group B) assessment: individual (n = 33)
CE	.28	.24	.57**
G&V	.19	-.10	.42*
Pr	.21	.18	.33
Total	.17	.06	.63**

Note. CE = Communicative Efficiency; G&V = Grammar & Vocabulary; Pr = Pronunciation. *p < .05, **p < .01.

With regard to the effect discussions between peers (RQ4), the correlations of Group A were low and not significant for any of the criteria (see Table 5), while the Communicative Efficiency and Grammar & Vocabulary criteria in Group B were significantly correlated with teacher assessment (r = .57 and .42). This contradicted expectations because it had been assumed that conducting assessments in pairs or groups would help reduce rater's subjectivity and increase the correlation between assessments by peers and the teacher. As indicated by Fukazawa's study (2009), a higher correlation can be obtained when multiple ratings of each performance are averaged. Therefore, a higher correlation may have resulted if the teacher asked each peer rater to decide on his or her own rating after discussion with a partner, rather than requiring them to decide on one rating together as a result of their discussion.

In both Groups A and B, the correlation between teacher and peer assessments of Communicative Efficiency was the highest of all three criteria (r = .24 and .57). This corresponds with the results obtained in Experiment 1. Thus, Communicative Efficiency was found to be the easiest criterion of the three for peers to assess.

In order to investigate how the two factors, assessor and criterion, were related, a two-way ANOVA with a 2 × 3 design (Rater [Peer, Teacher] × Criterion [Communicative Efficiency, Grammar & Vocabulary, Pronunciation]) was conducted for both Groups A and B. Data from Group A did not show a significant interaction between rater and criteria ($F(2,50) = 1.17$, p = .320, η^2 = .05) but the main effect of the

NII-Electronic Library Service

The Japan Language Testing Association

rater was significant ($F(1,25) = 6.08$, $p = .021$, $\eta^2 = .07$), thereby indicating that the scores given by peers were significantly higher than those given by the teacher. With regard to Group B, the results of the ANOVA did show an interaction between assessor and criterion. This indicates that peers and the teacher assigned different ratings to the three criteria. Therefore, *t*-tests were conducted as *post hoc* tests in order to compare the peer and teacher assessments for each criterion. The results revealed that peers had given higher scores than the teacher in Grammar & Vocabulary, Pronunciation, and total score: $t(32) = 5.31$, $p < .05$, $d = 1.00$; $t(32) = 3.25$, $p < .05$, $d = 0.67$; $t(32) = 6.22$, $p < .05$, $d = 0.94$, respectively. However, there was no significant difference between peer and teacher assessments in Communicative Efficiency: $t(32) = 1.85$, $p = .074$, $d = 0.31$. This again indicated that Communicative Efficiency was relatively easier to assess for peers than the other two criteria.

Next, in order to evaluate possible causes of and perspectives on the results reported thus far, the questionnaire was analyzed and the analysis results were summarized, as shown in Table 6. Addressing the importance of securing anonymity, Q1-5 asked students if they felt they could assess their peers more appropriately when the test taker did not know they were conducting the assessment. Here, the mean score was high for both groups at 4.65 and 4.36 on a 6-point scale. This result implies that anonymity is an important condition that can make peer assessment more reliable.

In addition, the students felt a sense of responsibility toward peer assessment, both when recording their speeches (4.35 and 4.09 in Q2-1) and assessing their peers (4.73 and 4.67 in Q2-2). These results correspond with what Saito (2008) referred to as "shared responsibility." As a washback effect, peer assessment may facilitate students' learning both when they are in the test taker and assessor roles.

However, peer assessment is somewhat demanding for students. The students responded that peer assessment was challenging (4.38 and 4.45 in Q3-5), and they could not rate the criteria with confidence (Q1-2 to Q1-4). In particular, they found it difficult to rate Grammar & Vocabulary (3.12 and 3.15 in Q1-3), which might explain the inconsistent correlations for that criterion across the three experiments (i.e., *r* ranges from -.11 to .42). Therefore, a teacher may need to explain to students which essential features of grammar and vocabulary they should focus on when they are conducting assessments.

With regard to the appropriateness of assessing in pairs or individually (i.e., Q1-6), the scores for both groups did not exceed 4.0. Specifically, the score for Group A, the members of which actually worked in pairs, was only slightly lower than that of Group B, the members of which worked alone (3.81 versus 3.88). This implies that many students did not particularly think that assessing in pairs was inappropriate. However, they might have compromised when rating with their peers and may not have assigned the rating that they believed was most accurate. In other words, sharing responsibility may have had negative effects. However, these reasons cannot be

NII-Electronic Library Service

The Japan Language Testing Association

verified and must remain speculative, since the difference in scores for Q1-6 was marginal and we could not find any other strong evidence to support this speculation in the questionnaire results.

Table 6

Survey on Peer Assessment in Groups A and B in Experiment 2

		Group A (n = 26)		*Group B (n = 33)*	
No.	Question items	*M*	*SD*	*M*	*SD*
1. When peer-assessing, I ...					
Q1-1	was actively engaged in the assessment.	4.15	0.88	3.97	0.92
Q1-2	was able to rate Communicative Efficiency with confidence.	3.38	1.10	3.24	0.94
Q1-3	was able to rate Grammar & Vocabulary with confidence.	3.12	0.91	3.15	1.03
Q1-4	was able to rate Pronunciation with confidence.	3.38	0.85	3.21	0.99
Q1-5	thought it was better for appropriate assessment if my name remained unknown to the test taker.	4.65	1.26	4.36	1.32
Q1-6	thought it was better for appropriate assessment to assess in pairs rather than independently.	3.81	1.06	3.88	1.36
2. Regarding peer-assessing, I ...					
Q2-1	thought I should work hard on it because I was also being assessed.	4.35	0.85	4.09	1.33
Q2-2	felt a sense of responsibility.	4.73	0.96	4.67	0.82
3. Regarding peer-assessment, I thought ...					
Q3-1	it was a good chance for me to review my retelling.	3.96	1.00	3.83	1.01
Q3-2	I could learn by listening to other student's retelling.	4.08	0.89	3.70	1.16
Q3-3	it was an important opportunity for assessing each other's performance.	3.92	1.13	3.76	1.00
Q3-4	we should have opportunities to assess each other's performance in regular classes.	3.27	1.19	3.52	1.00
Q3-5	it was difficult to assess other student's performance.	4.38	1.20	4.45	1.25

Finally, in order to compare the results of our study with previous studies, a meta-analysis was conducted (see Table 7) using *Comprehensive Meta-Analysis* (Borenstein, Hedges, Higgins, & Rothstein, 2005). For the meta-analysis, only the studies that reported correlations between single (not mean) scores from peer and teacher assessments were included. The conditions of these studies matched the definition of peer assessment used in our study. The overall results (Q = 73.16 with *df*

NII-Electronic Library Service

The Japan Language Testing Association

= 6, $p < .001$, $I^2 = 91.80$) showed that the studies were highly heterogeneous, with almost no consistency among their results (see Note 1 at the end).

Table 7

Meta-Analysis of the Literature with Results from Experiments 1 and 2

Study	[b]Language (Country/Region)	Prior Discussion	*r*	*N*	(95% CI)
Jafapur (1991)	L2 (Iran)	No	.44*	41	(0.15, 0.66)
Patri (2002)	L2(Hong Kong)	Yes	.85**	25	(0.69, 0.93)
Patri (2002)	L2(Hong Kong)	No	.49**	29	(0.15, 0.73)
AlFallay (2004) [a]	L2 (SaudiA)	Yes	.82**	200	(0.56, 0.73)
Experiment 1	L2 (Japan)	No	.17	63	(-0.08, 0.40)
Experiment 2 Group A	L2 (Japan)	Yes	.06	26	(-0.34, 0.44)
Experiment 2 Group B	L2 (Japan)	No	.63**	33	(0.37, 0.80)

Note. [a]The correlation coefficients for all conditions are combined. [b]The target language in all the studies is English. *$p < .05$, **$p < .01$.

Thus, in order to ascertain the cause of these inconsistent results, an analysis was conducted in order to determine how study results differed when prior discussion occurred. This analysis revealed that studies without prior discussion were homogeneous: $Q = 7.35$ with $df = 3$, $p = .062$, $I^2 = 59.19$, with an overall correlation coefficient of $r = .43$, $p < .001$. However, significant heterogeneity was found among studies that included prior discussion: $Q = 27.66$ with $df = 2$, $p < .001$, $I^2 = 92.77$. Compared to other studies including prior discussion, the correlation coefficient of the present study is remarkably low ($r = .04$). As mentioned earlier, one reason for the low correlation coefficient seems to be that the students exhibited reserved attitudes during discussion. These inconsistent results regarding the effect of discussion imply that the reliability of peer assessment varies according to classroom situations, such as relationships among peers and how peers assess each other. Therefore, the results might change if students are more comfortable engaging in discussion with one another or if individual scores from peer assessment following discussion are averaged.

Conclusion

The present study resulted in five main findings. First, students can reliably rate Communicative Efficiency because it does not demand the same degree of linguistic knowledge as Grammar & Vocabulary and Pronunciation. Second, peers tended to be lenient when rating the performances of fellow students. Third, peer and teacher assessments have their own distinctive patterns. For example, teachers rated Grammar & Vocabulary particularly severely, whereas peers rated this criterion more leniently.

NII-Electronic Library Service

The Japan Language Testing Association

This may be because the students were not sure of the correct grammar and vocabulary. Thus, on the basis of this information, RQ1 and RQ2 have been answered, as the findings confirm that peer assessments tend to correlate with or become similar to teacher assessment, particularly when the fluency of oral performance is the object of the assessment; on the other hand, peer assessments of linguistic aspects (i.e., Grammar & Vocabulary) tend to deviate from teacher assessment.

Fourth, the anonymity of raters was revealed to be an important factor in improving the reliability of peer assessment (RQ3). Although anonymity is difficult to preserve when peers are rating each other's oral performance, one possible measure is to exchange recordings with another class, as done in the present study.

Fifth, conducting assessment in pairs did not significantly improve reliability when compared to assessing independently (RQ4). Meta-analysis also revealed that the effect of peer discussion remained inconclusive, while studies on peer assessment without discussion were consistent, thereby revealing a moderate correlation between peer and teacher assessments ($r = .43$).

Overall, the correlations between peer and teacher ratings varied, thereby indicating that peer assessment cannot always be a reliable substitute for teacher assessment, particularly when single-score ratings are used. In addition, the diverse findings of the literature, including the present study, suggest that it is difficult to conclude which factors or conditions strongly affect the quality of peer assessment since the results are often vulnerable to contexts and circumstances in which peer assessments are administered. Therefore, it is necessary that greater attention be paid to the different roles peer assessment may play. In other words, instead of seeking a reliable assessment that closely resembles a teacher's assessment, the rating process may be more beneficial for collaborative learning if peer assessment is used for providing feedback rather than evaluating students.

Note 1. The Q value shows heterogeneity among multiple previous studies; a significant p value shows that the results of studies are completely different (Lipsay & Wilson, 2001). I^2 shows dispersion among effect sizes and is expressed as a ratio with a range of 0% to 100% (Borenstein, Hedges, Higgins & Rothstein, 2009).

Acknowledgments

We are very grateful to the anonymous reviewers for their thorough comments on an earlier version of this paper. This research was supported by Grant-in-Aid for Scientific Research (KAKENHI) (C) (23520744).

NII-Electronic Library Service

The Japan Language Testing Association

References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System: An International Journal of Educational Technology and Applied Linguistics, 32,* 407-425.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis (Version 2.2.023) [Computer software]. Englewood Cliffs, NJ: Biostat.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis.* London: John Wiley & Sons, Ltd.

Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus self assessment of oral business presentation performance. *Business Communication Quarterly, 64,* 23-42.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20,* 289-300.

Fukazawa, M. (2009). Speech ni okeru seito-sougo-hyoka no datousei: Koumoku-outou-riron wo mochiite [The validity of peer assessment in speech: Using item response theory]. *STEP Bulletin, 21,* 31-47.

Hirai, A., & Koizumi, R. (2008). Validation of an EBB scale: A case of the story retelling speaking test. *JLTA (Japan Language Testing Association) Journal, 11,* 1-20.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6,* 151-167.

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18,* 379-385.

Jafapur, A. (1991). Can naïve EFL learners estimate their own proficiency? *Evaluation and Research in Education, 5,* 145-157.

Koizumi, R. (2005). Predicting speaking ability from vocabulary knowledge. *JLTA (Japan Language Testing Association) Journal, 7,* 1-20.

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education, 33,* 179-190.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Miller, L., & Ng, R. (1994). Peer assessment of oral language proficiency. Retrieved from http://sunzi.lib.hku.hk/hkjo/view/10/1000076.pdf/

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21,* 239-250.

NII-Electronic Library Service

The Japan Language Testing Association

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19,* 109-131.

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

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

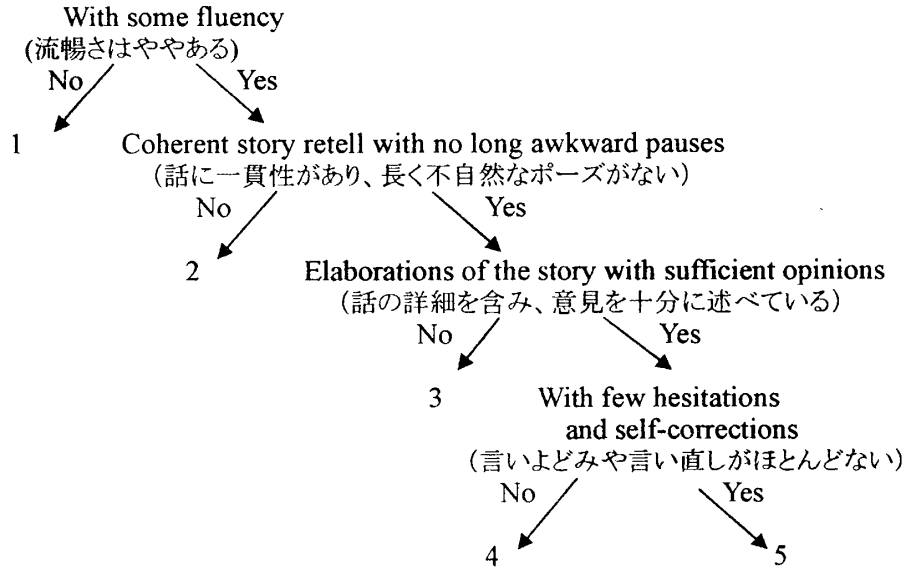
Yamashita, Y. (2008). *The effect of oral output on Japanese EFL learners' retelling.* Unpublished graduation thesis, University of Tsukuba, Japan.

NII-Electronic Library Service

The Japan Language Testing Association

Appendix. *The Criteria in EBB Scale Used in the Experiments*

1. Communicative efficiency(伝達能力)



定義・目安

- Communicative efficiency の中には、coherence, fluency, volume, content の観点が入っている。
- coherent story: so, then,など物語の流れを正しくつかみやすく話している。

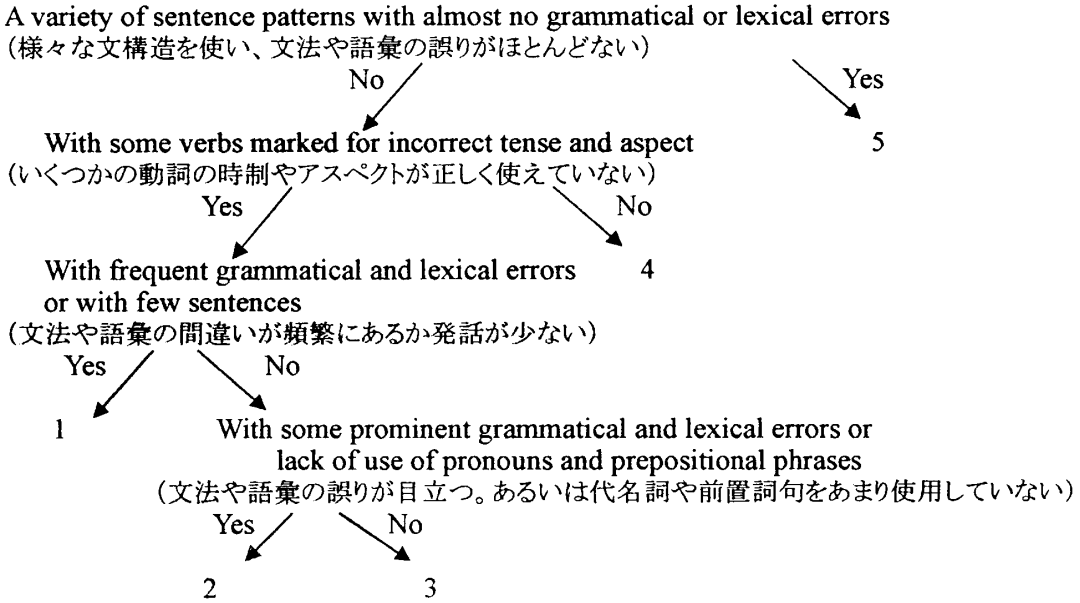
基本的に時間の流れ順に言っていて、流れがとりやすいものは、so などがなくても coherent と判断してよい。

- Elaboration of the story: 最低限の物語の骨格 (key storylines) に加え、詳しく述べている。
- sufficient opinions: ストーリーに関する自分の意見を (最低 3 文以上で)、適切に述べている。
- long awkward pauses : 約 4、5 秒以上の不自然なポーズ。

retelling と opinion の間のポーズなど、4, 5 秒以上あっても不自然でなければ、long と考えない

- fluency (speed, pause, hesitation を含めた流暢さ。単語ごとに発音せず、決まり文句はまとめて話す、frequent long pauses がない、frequent hesitation がない)
- few hesitations : 話すペースが一貫しており、内容を理解するのに妨げになるほどの hesitation がない。

2. Grammar and vocabulary(文法と語彙)



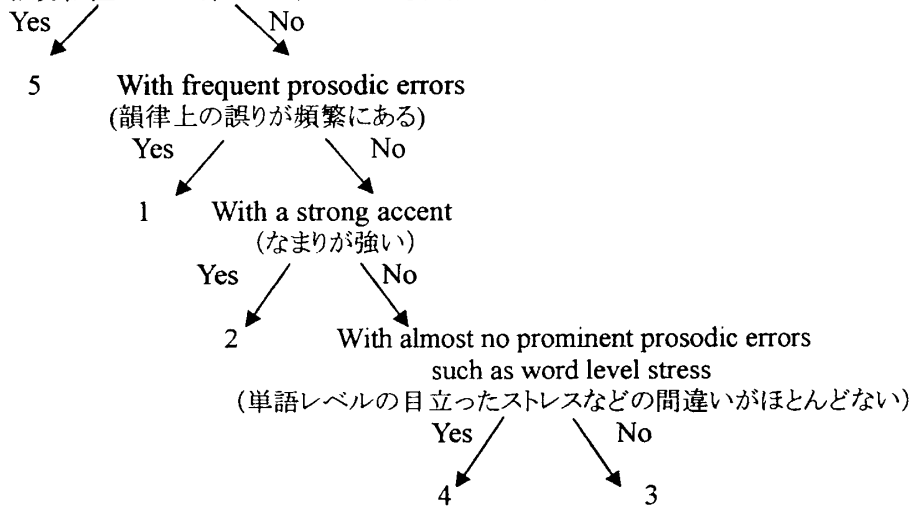
NII-Electronic Library Service

定義・目安

■ Grammar and vocabulary の中には、accuracy, complexity の観点が入っている。
■ With some verbs marked for incorrect tense and aspect：時制やアスペクト（進行形、完了形など）の間違いはあっても、あまり多くない。1, 2 個以内。その間違いがほとんど気にならない。
■ Use of pronouns and prepositional phrases
e.g., 1. They (←Bob and Jean) didn't want to leave the beach.（但し、原文で pronoun を使っていないのであれば厳しくつけない）
2. I want to go with him [prepositional phrase].
■ error の判定は、dysfluency markers を除いた形で考える。
例：She {like} likes English. Correct
■ lexical error: soonerly など、英語として存在しない語
■ Tense and aspect: major error と考えた
Pronouns, prepositional phrases: minor error と考えた（また、習得上、正しい時制とアスペクトを使用できるほうが難しいかどうかを確認する）
■Grammar で、発話が少ないことで 1 になる理由は、発話が少ないということは
learners don't have sufficient grammatical knowledge even to construct short sentences だから。
■ With few sentences: 文の数については、opinion を含めて数える。and 等で続けている場合は、分けて 2 文にカウントする。4、5 文以下の発話は few sentences と考える。
■ some prominent grammatical and lexical errors は、意味が伝わりにくい目立った誤りが 2 個程度ある。
■ frequent grammatical and lexical errors は、発話量に比べて、比較的誤りが多い場合。

3. Pronunciation (発音)

Accurate pronunciation with correct stress and natural intonation
（正確な発音でかつ強勢位置が正しく、イントネーションも自然である）



定義・目安

■ Pronunciation (includes stress, accent, intonation)
■ prominent prosodic errors (韻律の誤り。容易にわかる発音やストレスの位置の間違いやもごもご話していて、発音がはっきりしない。例：Florida の発音 ri にアクセントを置くなど)。
(聞き分けにくい細かな点はあまり厳しくつけない。例：l / r, sea / she, found / hound で、微妙なものなど。ただし、はっきり分かる発音の誤りは、l / r, sea / she, found / hound のようなものであっても誤りとする)
■ Jean を Jane と言うなど、名前の発音誤りは誤りと考えない。
■ なまりは、日本人なまり、韓国人なまりなど、日本人のなまりには限定しない。
■ 早く判断できるものを上にした。

NII-Electronic Library Service