Rater Reliability in Peer Speech Assessment done by CEFR A2 EFL Learners

Seiji Takahashi
Baiko Gakuin University

Abstract

This paper aims to examine to what extent peer speaking assessment, conducted by CEFR A2 EFL learners, aligns with their instructor's assessment. It also investigates whether there are differences in the assessment approach between two learner subgroups with varying English proficiency within the same CEFR band. The participants in this study were 21 EFL university students, most of whom had English proficiency at the CEFR A2 level. These students were enrolled in a 15-lesson ESP course that focused on presentation and paragraph writing. Each participant's presentation was evaluated by both the author and the other participants based on seven criteria, encompassing linguistic and non-linguistic qualities. The obtained scores were then compared. The results indicated that peer assessment of linguistic qualities showed a weak correlation with the instructor's assessment, whereas assessment of non-linguistic qualities exhibited a medium-to-strong correlation. These findings suggest that CEFR A2 EFL learners may find it challenging to evaluate their peers' speech as accurately as their instructor. Additionally, the relative English proficiency within the same CEFR band group had a minimal impact on achieving a more precise rating in speaking.

Keywords：Peer assessment　Speech evaluation　Grading

Background

In second language (L2) pedagogy, peer assessment (PA) generally refers to the evaluation of language learning products by other learners with similar linguistic skills or knowledge (Brown, 1998; Matsuzawa, 2002; Topping, et al., 2000). PA has potential benefits, including

practical advantages such as reducing teachers' workload (Brown, 1998). Educators also argue that it can help clarify students' goals (Fukazawa, 2009), promote autonomy (Okuda & Otsu, 2010), and foster a sense of responsibility towards others' learning (Brown, 1998). PA is particularly valuable in reducing teachers' bias in speaking evaluation, which tends to be subjective (Fulcher, 2003; McNamara, 1996).

Despite the potential benefits of PA for learners' performance, one barrier to its implementation in the classroom is the contradictory findings regarding its reliability in previous studies. While some researchers have found a strong-to-medium correlation between PA and instructor assessments (e.g., Fukazawa, 2009; Okuda & Otsu, 2010), others (e.g., Freeman, 1995; Kasamaki, 2020; Orsmond, et al., 1996) have reported its lack of reliability. In short, scholars hold divided opinions on its practicality. One possible reason for these conflicting results may be attributed to learner variables. Miller and Ng (1996) suggested that the reliability of PA depends on the evaluators' linguistic ability. Similarly, Luoma (2004) argued that learners are not qualified to assess the linguistic aspects of their peers due to their limited second language proficiency. Speaking assessment, in particular, is considered the most challenging aspect of language evaluation (Fulcher, 2003), as assessors are required to evaluate both grammatical correctness (e.g., grammatical structure, lexicon, collocation) and pronunciation accuracy (e.g., enunciation, prosody). Moreover, the assessment often extends to nonverbal communication, such as gestures, posture, and eye contact. Therefore, learners' linguistic proficiency is considered crucial in PA for evaluating speaking performance.

In existing studies on speaking assessment, only a limited number of researchers have provided details about participants' proficiency levels, such as CEFR bands or scores/grades from an English proficiency test. One of these studies that specified participants' proficiency is Shimura (2006), in which university students in Japan were divided into three groups (high, middle, and low) based on their TOEFL scores. The study aimed to explore how the English proficiency of each group affected the accuracy of their peer assessment of presentations. The results indicated that the middle group exhibited the strongest correlation with the benchmark scores (scores obtained from assessments conducted by teachers), while the other groups also showed a middle-to-strong correlation with the benchmark. Therefore, the researcher concluded that higher proficiency does not necessarily lead to more accurate peer assessment. Another relevant study is Kasamaki (2020), which divided first-year university students in Japan into three groups based on their TOEIC IP

scores. The study examined the extent to which participants' assessments aligned with their instructors' assessments and whether there were differences in peer assessment among the groups. The findings revealed a medium correlation between participants' assessments in each group and the baseline assessment. As a result, the researcher suggested that peer assessment was not reliable enough, and no significant differences were observed among the groups.

The studies have revealed that English proficiency is not a reliable predictor for accurate rating, and peer assessment is not as reliable as instructors' evaluations. However, further research is needed to determine the extent to which peer assessment conducted by learners in other proficiency groups can be considered reliable. In Shimura's study (2006), the participants in the highest group had an advanced level of English proficiency for Japanese university students, with TOEFL scores higher than 550, which corresponds to CEFR levels B1 to B2 (ETS, 2022a). In the case of Kasamaki's study (2020), the learners were considered to have CEFR levels A2 to B1, as their TOEIC IP scores ranged from 290 to 680, with a mean score of 464, which is average or above average for first-year university students in Japan. It is worth noting that the average score of the same test for university students (first to fourth-year students) is 471 (ETS, 2022b). These findings do not fully represent the entire population of EFL learners in Japan, indicating the need for more data to fill the paucity.

To that end, this study aims to examine the reliability of peer assessment of speech conducted by a learner group with lower proficiency compared to the participants in previous research. Additionally, the study aims to explore whether there are differences in assessments made by learners in two proficiency subgroups, thereby investigating how learners' relative proficiency influences their assessments. To address these goals, the following research questions were formulated:

1. To what extent does peer assessment of English speech, conducted by English learners with CEFR A2 proficiency, align with their instructor's assessment?

2. Are there any differences in peer speech assessments between two proficiency subgroups within the same CEFR band? If so, what are these differences?

Methods

Participants

A total of 21 first-year students (6 males, 15 females) enrolled in English and international business programs at a private university in Japan were recruited as participants for this study. They were taking a 15-lesson ESP course aimed at developing presentation and writing skills in English. All participants were native Japanese speakers who learn English as an L2. The participants' English proficiency was assessed using the score of TOEIC IP test, which evaluates listening skills and grammatical knowledge, both of which are considered important qualities for speech assessors (e.g., Buck, 2001). Their TOEIC IP scores ranged from 165 to 570, with a mean of 409 and a standard deviation of 98.8. With the exception of one participant, all others fell within the CEFR A2 level (ETS, 2022a), which signifies basic users of the target language according to the Council of Europe (2001).[1]

For the second research question, the participants were divided into two groups based on their English proficiency. Despite being in the same class and CEFR band, the participants' proficiency levels varied, as evident from the range of TOEIC IP scores and standard deviation. The standardized variate (SV) was calculated for each participant using their TOEIC IP scores, and those with an SV above 0.00 were classified into the upper group (n=9), while the remaining participants were placed in the lower group (n=12). In the end, a total of 11 students comprised the lower group, as one student with missing values during data analysis was excluded.

Data Collection

Presentation

The data collection for the current study took place during the third quarter of 2022, specifically in October and November. Within one of the 15 lessons of the course, which followed the Speaking of Speech, Premium Edition textbook (LeBeau, 2009), each student delivered a three-minute presentation recommending a product to their peers. The focus of the presentation was solely on the introduction part, which corresponded to the immediate unit covered in the textbook. The learners were instructed to adhere to the introduction structure taught in the textbook, which included a greeting, a title, a hook, and an overview of the entire presentation. After listening to each presentation, each student assessed all the

other students (excluding themselves) using an assessment criteria list. This resulted in a total of 20 presentations being evaluated by each student. The researcher also assessed all the presenters (i.e., a total of 21 presentations). The assessments by both the students and the researcher were conducted simultaneously.

Assessment Criteria

The assessment criteria list consists of seven points, namely: 1. Pronunciation, 2. Expression, 3. Content, 4. Posture and Eye Contact, 5. Gesture, 6. Voice, and 7. Slide. The last criterion, Slide, was labeled as "PPT" on the printed list used (see Appendix). Each criterion was measured on a 5-point scale, with 1 representing the lowest score and 5 representing the highest score. Prior to the presentations, the participants received an explanation from the researcher regarding what to consider when assessing each criterion. Regarding Pronunciation, the participants were instructed to assign a score of five if they found the presenter's English to be sufficiently intelligible, with moderate accuracy in stress and intonation, even if there were some influences from their L1 (Japanese) on individual sounds. In terms of Expression, which pertained to grammatical aspects, the participants were instructed to assign the highest score if the presenter demonstrated accurate use of basic grammatical structures, even if errors occurred when attempting to use more complex grammatical forms. For the Content criterion, the participants were instructed to assign a score of five if the presentation covered all the required elements (i.e., greeting, title, hook, and overview) and if the hook part was ingenious and appealing. In the case of Posture and Eye Contact, Gesture, and Voice, the participants had previously learned how to effectively utilize these physical aspects in speech during earlier lessons in the course. Additionally, they were instructed to consider the model presentations they had watched on DVD as a benchmark for assigning a score of five in those criteria. However, the Voice score was excluded from the statistical analysis because the perception of how loud a presenter's voice was may have been influenced by the assessors' seating positions within a relatively large classroom given the number of participants.

Due to the need to adhere to a shared syllabus with other course instructors, the participants did not have the opportunity for rating training, which is commonly recommended to enhance the reliability of assessments (Fulcher, 2003; McNamara, 1996; Weir, 1990). Nonetheless, it can be assumed that the participants had a good understanding of what constitutes a strong presentation, as they had repeatedly viewed model presentations

through video materials and learned essential elements for delivering an effective presentation (e.g., persuasive speaking style, proper posture, effective gestures, and visually appealing slides) during the course. The course syllabus explicitly stated that peer evaluation would contribute to the participants' final grades, and they were instructed to engage in the assessment process seriously. Additionally, they were encouraged not to hesitate in assigning lower scores to their peers, as their evaluations would remain anonymous to other students.

Benchmark Scores

Assuming the instructor's assessment to be accurate, the scores provided by the researcher (referred to as Teacher's Assessment, or TA) were utilized as a reference point to investigate the reliability of peer assessment. In reliability studies, employing multiple assessors is recommended to mitigate potential score variations. Furthermore, if an evaluation is conducted by a single assessor, it is crucial to consider intra-rater reliability to ensure a rigorous rating process (Fulcher, 2003). However, as one of the objectives of this study was to identify a more effective approach for evaluating students' speech within a real classroom setting, employing more than one assessor was deemed unrealistic. Additionally, reassessing the same presentation at a later time interval to establish intra-rater reliability was considered impractical in everyday educational practice. Hence, each student's presentation was assessed only once by the researcher to establish the benchmark score.

Data Analysis 1: The Degree of Agreement between TA and PA

The scores assigned by the participants were initially calculated to determine the mean sectional scores and overall scores for each presenter using Microsoft Excel. Specifically, the mean sectional score for each presenter was obtained by summing all the scores received for each criterion given by the 20 participants, and then dividing the sum by 20. To assess the strength of the relationship between the Teacher's Assessment and the mean Peer Assessment scores, Spearman's rank-order correlation coefficient was employed. The statistical analysis was conducted using IBM SPSS Statistics (version 29). The choice of Spearman's test was made as the data set did not appear to follow a normal distribution, as determined through a visual examination of the scatterplot (Larson-Hall, 2015, p. 477).[2]

Data Analysis 2: Difference between PA from Two Proficiency Groups

The purpose of this analysis was to compare the scores provided by participants in the upper

and lower proficiency subgroups to investigate potential differences in their assessments of the presenters. Specifically, the focus was on the scores given by each participant to the other participants, rather than the scores received by each participant from their peers during the initial evaluation. To clarify this point, Spearman's rank-order correlation coefficient was calculated between the sectional scores (e.g., scores for Pronunciation) given by each participant to the 20 presenters and the scores assigned by the researcher to the 21 presenters. The average correlation coefficients for each section were then determined for both subgroups. Subsequently, a Mann-Whitney U test was performed to determine whether there were significant differences between the coefficients of the two subgroups.

Results

Analysis 1 aims to assess the level of agreement between the scores assigned to each student by the researcher and the scores provided by the participants. To gain an overview of both the TA and PA, the mean scores and standard deviations were calculated for each criterion as well as the total scores, as shown in Table 1. The analysis revealed that the PA scores were considerably higher than the TA scores. Additionally, the standard deviations of the PA scores indicated that the participants utilized a significantly narrower range of marks compared to the researcher.

Table 1. Descriptive Statistics for TA and PA by Individual Criterion

	TA		PA	
	M	*SD*	*M*	*SD*
1. Pronunciation	3.95	0.67	4.64	0.42
2. Expression	3.90	0.77	4.64	0.22
3. Content	3.86	0.85	4.64	0.21
4. Posture/EC	3.33	1.20	4.41	0.33
5. Gesture	3.14	1.49	4.25	0.44
6. Slide	3.67	0.66	4.75	0.11
Total	21.85	4.39	27.33	1.47

Note. EC indicates Eye Contact.

The results of Spearman's rank-order correlation coefficient are presented in Table 2. The findings indicate that PA scores for Posture and Eye Contact (r = .78) and Gesture (r = .85)

exhibited a strong association with the TA. However, scores for Pronunciation (r = .32) and Expression (r = .20) displayed a weak correlation. Content (r = .52) demonstrated a moderate correlation. On the other hand, Slide (r = .17) exhibited the weakest correlation among all the criteria, despite having the highest mean score and the narrowest standard deviation. Notably, the Total score showed a strong correlation, even though half of the criteria displayed weak correlations.

Table 2. Correlation between TA and PA

	Correlation Coefficient (r)
1. Pronunciation	.319
2. Expression	.200
3. Content	.522*
4. Posture/EC	.799**
5. Gesture	.853**
6. Slide	.172
Total	.788**

Note. * and ** indicate statistical significance at $p < .05$ and $p < .01$ respectively. df = 19

To conduct Analysis 2 and compare the two subgroups, Table 3 summarizes the minimum, maximum, and mean correlation coefficients for each section in the upper and lower groups. In terms of Pronunciation, both groups exhibited weak correlations with TA (r = .348 and .243, respectively). However, the lower group displayed a wider range of scores for Pronunciation (SD = .305) and demonstrated a stronger negative correlation (r = -.463) as well as a stronger positive correlation (r = .714). Conversely, the upper group had the narrowest range of scores (SD = .155) among all criteria. Regarding Expression, both groups showed the same weak correlation (r = .193). No significant associations were observed between TA and the PAs of either group for Content and Slide. In relation to Posture and Eye Contact, the upper group displayed a weak correlation (r = .266), while the lower group exhibited a moderate correlation (r = .450). For Gesture, both groups showed a medium correlation (r = .452 and .490). The Total scores of both groups displayed a medium correlation with TA (r = .413 and .489).

Table 3. Descriptive Statistics for Correlation between TA and PA-Upper, and TA and PA-Lower

	PA-Upper				PA-Lower			
	M	*SD*	*Min.*	*Max.*	*M*	*SD*	*Min.*	*Max.*
1. Pronunciation	.348	.155	.050	.475	.243	.305	-.463	.714
2. Expression	.193	.285	-.318	.441	.193	.320	-.450	.635
3. Content	.061	.219	-.265	.274	-.014	.308	-.428	.300
4. Posture/EC	.266	.282	-.308	.585	.450	.215	.110	.779
5. Gesture	.452	.295	-.286	.671	.490	.257	-.074	.866
6. Slide	-.062	.232	-.461	.234	.076	.255	-.396	.399
Total	.413	.230	.055	.684	.489	.231	.025	.855

To examine the potential differences in assessments between the two groups, a Mann–Whitney U test was conducted. Consistent with the previously mentioned results, no significant difference was observed between the two groups (see Table 4). Furthermore, effect size was calculated, revealing a small effect size (r = .104) for Posture and Eye Contact, while the effect sizes for the remaining criteria were negligible.

Table 4. Results of Mann–Whitney U test (n = 20)

	U	*Z*	*p*	*Effect size (r)*
1. Pronunciation	32.0	-.993	.351	.054
2. Expression	40.0	.136	.930	.001
3. Content	24.0	-.122	.953	.001
4. Posture/EC	68.0	1.406	.175	.104
5. Gesture	49.0	-.038	1.000	.000
6. Slide	36.0	1.061	.328	.080
Total	57.0	.570	.603	.017

Discussion and Conclusion

Within the context of 21 Japanese first-year university students whose English proficiency ranged from CEFR A2 to barely B1, this study aimed to investigate the level of agreement between their assessments and those of their instructor. Additionally, the study explored whether learners in two proficiency subgroups evaluated their peers' presentations differently. Regarding the first research question, the results indicated significant variation

in the degree of agreement across different criteria. Significant correlations were observed between the benchmark scores and the nonverbal criteria of Posture and Eye Contact (r = .799) as well as Gesture (r = .853), in line with expectations. These findings align with the inherent nature of these criteria, which rely on nonverbal communication and do not require linguistic proficiency for assessment. The medium correlation observed in Content (r = .522) aligns with the anticipated findings, as this criterion is not directly contingent on L2 linguistic ability. However, it is important to note that certain aspects of L2 linguistic proficiency, such as listening skills and vocabulary knowledge, are deemed indispensable for comprehending the content of presentations delivered in the L2. Conversely, Pronunciation and Expression showed weak correlations, likely due to the difficulty learners face in evaluating linguistic performance. Surprisingly, Slide displayed the weakest correlation among all the criteria, despite being a non-linguistic aspect. The high mean PA score for Slide and its narrow standard deviation suggest that many participants may have given a score of five if they perceived a presenter's slide to be visually appealing. While the instructor made efforts to differentiate the quality of slides, participants may have exerted minimal effort in doing so. Finally, Total scores were strongly correlated with TA. This indicates that the PA scores for non-linguistic criteria, which demonstrated medium-to-strong correlations with TA, compensated for the weaker correlations observed in other areas. In summary, it is prudent to refrain from concluding that PA was as reliable as TA; instead, the overall reliability of PA appears to be questionable.

Regarding the second research question, no significant difference in rating behaviors was observed between the two groups. While the upper group demonstrated a stronger correlation with TA in terms of Pronunciation, and their standard deviation was considerably smaller than that of the lower group, suggesting their potential as more competent assessors, overall, the participants' relative English proficiency had only a marginal impact on how they evaluated their peers' speech. In brief, regardless of their English proficiency levels, the participants exhibited similar assessment patterns when evaluating their peers.

The above-mentioned results partially support existing literature indicating that PA generally yields higher scores than TA, and the participants tend to use a narrower range of marks compared to the instructor (e.g., Freeman, 1995; Okuda & Otsu, 2010; Kasamaki, 2018). However, the overall reliability of PA in the current study contradicts previous research (Okuda & Otsu, 2010; Fukazawa, 2009), where learners' assessments were strongly aligned with those of their instructors. One possible explanation for this discrepancy is that

the participants in the earlier studies had higher English proficiency levels compared to the learners in the current study. Specifically, while Okuda and Otsu (2004) did not specify the English proficiency of their participants, they were students from a national university, which typically signifies a leading institution in the Japanese context. Fukazawa (2009) did not explicitly specify the English test scores or CEFR levels of the participants either, but it was mentioned that they hailed from a prestigious high school with a strong emphasis on English and science. This implication strongly suggests that the participants possessed English proficiency levels above the average for their age group.

Another important point to consider is that the participants in the current study did not receive any rating training, whereas Fukazawa (2009) and Okuda and Otsu (2010) provided rigorous training sessions prior to the assessment. Although the participants in the present study were exposed to model presentations and received explanations on how to evaluate a presentation, they did not engage in actual rating practice. Furthermore, since the participants were only exposed to excellent model presentations, they lacked experience in listening to poorer speeches, which may have been crucial for their development as better assessors, particularly considering that most of the presentations they evaluated in the course did not reach the same level as the models.

Regarding the second research question, the subgroups showed no difference in their assessments. This finding contradicts Shimura's (2006) study, likely because the participants in her study had a wider range of English proficiency. The highest group in her study consisted of participants classified as CEFR B1 or B2 (i.e., independent users), while the lowest group comprised CEFR A2 (i.e., basic users). Thus, there was a clear distinction in English proficiency between the two groups. In contrast, in the current study, the difference in English proficiency between the two groups, or the gap between the lowest and highest proficiency levels, was narrower since all participants, except one, fell within CEFR A2. Consequently, their relative English proficiency, which is thought to be related to the quality of evaluation, may have made only a marginal contribution to their assessments.

As the present study highlights the reliability of peer assessment, the findings have pedagogical implications. The results suggest that learners' assessment of most non-linguistic qualities shows medium-to-strong associations with the benchmark. Therefore, teachers could consider incorporating these scores into grading, particularly when evaluating the content of a presentation, as it tends to be subjective when assessed by a single individual (Fulcher, 2003; McNamara, 1996). Additionally, educators should consider

integrating peer assessment activities into lessons, even if the evaluations are not used for grading due to their unreliability. Engaging in peer assessment can help develop learners' critical perspectives on others' performance, which, in turn, can contribute to their own improvement. Moreover, it can aid in clarifying learners' goals (Fukazawa, 2009), fostering autonomous learners (Okuda & Otsu, 2010), and cultivating a sense of responsibility towards others' growth (Brown, 1998).

To conclude, several limitations of this research should be acknowledged for future studies. First, neither intra-rater nor inter-rater reliability, which are known to increase the reliability of benchmark scores and thus the results, were obtained in the present study. As previously mentioned, this decision was intentional to reflect real-world situations where a single instructor typically assesses learners only once. However, it may have affected the quality of comparisons between the present study and previous research, where either inter- or intra-rater reliability was obtained. This also applies to the lack of rater training in the present study. Second, the criteria used in this study differed from those in previous studies, making it challenging to compare the results directly. Third, the three-minute duration of the presentations might not have provided sufficient speaking data for assessors to accurately evaluate the seven criteria. The length of the speeches and the number of criteria should be reconsidered in future studies. Moreover, some students may have been hesitant to assess their peers critically (Hanrahan & Isaacs, 2001). Such sociopsychological factors in the classroom were not considered in the present research, but they could have influenced the accuracy of the assessments and, consequently, the results of the correlation coefficient. Finally, the research was based on data from only 21 participants, limiting the generalizability of the findings to other CEFR A2 EFL learners. To investigate the reliability of peer assessment among CEFR A2 learners, future research should gather data from a larger number of participants.

Notes

1. ETS (2022) provides separate CEFR bands for the listening and reading sections, allowing learners to fall into different bands for each section (e.g., A2 for listening and B1 for reading). However, in the present study, the total score of both sections was used to determine learners' overall CEFR band. Specifically, learners with a total score of more than 550 were categorized as B1, while those with a total score below that threshold were referred to as A2, regardless of their sectional scores.
2. Larson-Hall (2015, p. 477) recommends using visual judgment of the scatterplot to assess normality, as the Kolmogorov-Smirnov test may not be sufficiently sensitive when the dataset is

small, making it difficult to determine normality based solely on the test results.

3. Takeuchi and Mizumoto (2012, pp. 125-126) suggest that the statistical significance of findings is heavily influenced by the number of participants involved. For instance, in a group of 17 participants, a correlation coefficient would need to be as high as $r = .412$ to achieve statistical significance at $p < .05$. However, if the number of participants increases to 102, statistical significance at $p < .05$ can be obtained even with a correlation coefficient as low as $r = .164$. Therefore, the researchers recommend considering the correlation coefficient itself, regardless of statistical significance, when discussing the strength of correlations.

References

Brown, J.D. (Ed.). (1998). *New ways of classroom assessment*. Alexandria, VA: Teachers of English to Speakers of Other Languages

Buck G. (2001). Assessing Listening. New York, NY: Cambridge University Press.

Council of Europe (2001). Common European Framework of Reference for Languages; Learning, teaching, assessment.

ETS (2022a). TOEFL score and CEFR level. Retrieved from https://www.ets.org/toefl/itp/scoring.html. January 2023.

ETS (2022b). TOEIC Program DATA & ANALYSIS 2022. Retrieved from https://www.iibc-global.org/library/default/toeic/official_data/pdf/DAA.pdf. January 2023.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, 20(3), 289–300.

Fukazawa, M. [深澤真] (2009) Speech ni okeru sougohyouka no datousei – koumoku outou riron wo mochii te [スピーチにおける生徒相互評価の妥当性—項目応答理論を用いて—: Validity of peer assessment of speech performance using item response theory]. *Step Bulletin*, 21, 31–47.

Fulcher, G. (2003). *Testing Second Language Speaking*. Essex: Pearson Education.

Hanarahan, J., S., & Issacs, G. (2001). Assessing self- and peer-assessment: the students' views. Higher Education Research & Development, 20(1), 53–70.

Kasamaki, T. [笠巻知子] (2018). Gakusei no presentation ryoku ha hyoukasya toshiteno gakusei no hyoukaryoku ni eikyou wo oyobosuka? [学生の「プレゼンテーション力」は，評価者としての学生の評価力に影響を及ぼすか?: Is a student's skills as a presenter related to his/her skills as a rater?]. *Language Education & Technology*, 55, 97-122.

Kasamaki, T. [笠巻知子] (2020). Gakusei ni yoru sougohyouka ryoku to presentation ryoku ni oyobosu youin ni kansuru jissyou kenkyu – gakusei no sougo hyouka to kyouin ni yoru hyouka to no soukan to hyoukasya training ni motozuite [学生による相互評価力とプレゼンテーション力に及ぼす要因に関する実証的研究 —学生の相互評価と教員による評価との相関と評価者トレーニングに基づいて—: Empirical research on factors affecting peer assessment and presentation – Based on the relationship between assessment done by instructors and learners, and on rater training]. [Unpublished doctoral dissertation].

Larson-Hall, J. (2015). A guide to doing statistics in second language research using SPSS and R. New York, NY: Routledge.

LeBeau, C. (2020). *Speeking of Speech Premium Edition*. National Geographic Learning.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

Matsuzawa,

McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.

Miller, L., & Ng, R. (1996). Autonomy in the classroom: Peer assessment. In R. Pemberton, E.S.L. Li, W. W.F. Or, & H.D. Pierson (Eds.), *Taking control: Autonomy in language learning* (pp.133-146). Hong Kong: Hong Kong University Press.

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

Okuda, R., & Otsu, R. (2010). Peer assessment for speeches as an aid to teacher grading. *The Language Teacher, 34* (4), 41–47.

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.

Shimura, M. (2006). Peer and instructor assessment of oral presentations in Japanese University EFL classrooms: A pilot study. *Waseda Global Forum 3*, 99-107.

Takeuchi, O., & Mizumoto, A. [竹内理 & 水本篤] (2014). *Gaikokugo Kenkyu Handbook – Yori yoi kenkyu no tame ni* [外国語教育研究ハンドブック―研究手法のより良い理解のために: Handbook for Foreign language Education Research – Improving Research Methodologies]. Tokyo, Japan. Shohakusha [松柏社].

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of aca demic writing between postgraduate students. *Assessment & Evaluation in Higher Education*, 25(2), 149-169.

Weir, C.J. (1990). *Communicative language testing*. New York: Prentice Hall.

Appendix
Performance Evaluation Sheet

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

Presenter ____________________

Pronunciation	1	2	3	4	5
Expressions	1	2	3	4	5
Content	1	2	3	4	5
Posture/Eye contact	1	2	3	4	5
Gestures	1	2	3	4	5
Voice	1	2	3	4	5
PPT	1	2	3	4	5

Total [ ]

ID ____________________ **Name** ____________________