Peer assessment for speeches as an aid to teacher grading

Rieko Okuda
Hiroshima University / Hiroshima Shudo University

Rika Otsu
California State University

Keywords
peer-assessment (PA), teacher assessments (TA), grading

In this study, we examined the level of agreement between teacher assessment and peer assessment during a speech presentation in an EFL context. A total of 88 students assessed speeches delivered by their peers. After four practice rounds of evaluating each other in small groups, a final assessment, including teacher assessment was conducted on speeches delivered to the whole class. Before each assessment, specifics on how to conduct the evaluations were explained by an instructor through visual demonstrations. A strong correlation (r = .82) was found between teacher marking and peer marking which indicates the viability of incorporating peer assessment into students' final scores when proper guidance is provided. A questionnaire administered after the final speech revealed that most of the students had found peer assessment useful.

本研究では、学生によるスピーチについての教員評価（TA）とピア評価（PA）の一致の度合いを調べた。被験者88人は、小グループ内でスピーチとPAを４回実施した後、クラス全員の前でスピーチを行った。この最終スピーチではTAとPAを同時に実施した。評価基準については、教員が実演を交えて項目ごとに説明し、それをPA実施のたびに繰り返した。その結果、TAとPAの間には高い相関（r = .82）が得られ、最終評価へのPA組み入れが可能であることが示唆された。またPA実施後のアンケート調査の結果から、多くの学生がピア評価活動を「有益である」と評価しているのが分かった。

Learner autonomy in the classroom has been increasingly emphasized in recent years and as a result there has been a greater focus on both self-assessment and peer-assessment as educational tools (Brown, 1998; Clifford, 1999; Miller & Ng, 1996). Peer assessment (PA), which can be defined as "an arrangement for peers to consider the level, value, worth, quality or successfulness of the products or outcomes of learning of others of similar status" (Topping, Smith, Swanson & Elliot, 2000, p. 150), appears to affect student motivation while reducing some of the rating responsibilities of teachers. In a speech presentation class for example "giving students the opportunity to evaluate their peers" (Brown, 1998, p. 67) on skills such as speaking at an appropriate volume and rate, enunciating clearly, or making good eye contact "not only gives them an important sense of responsibility for their fellow students' progress, but also forces them to concentrate on the skills during their own presentations" (p. 67). In addition to this motivational effect, PA is believed to reduce the teachers' marking burden. Boud (1989) argues that "if there is a high correlation between marks generated by students and those generated by staff … there is potential for saving of staff time on the often tedious task of marking" (p. 22). That is, if students can accurately assess their peers "teacher assessment could be supplemented with peer-assessment" (Patri, 2002, p. 125). If this is the case PA may help teachers in testing their students' oral skills.

One possibility is that teachers can use PA as a supplement to teacher assessment (TA) for speeches delivered to the whole class. Weir (1990) claims that the assessment of spoken language is potentially "problematic .… given that no recording of the performance is usually made" (p. 80). Speech sound disappears immediately after it is produced and cannot be repeated. Therefore, assessments in oral tests have to be made "either while the performance is being elicited or shortly afterwards" (p. 80) and grades cannot be reconsidered as many times as necessary. However, because a student's speaking skill in the classroom is usually assessed by a single teacher, teachers are required to stay attentive throughout. This need for constant attention is often tiring and teachers may drift off at times especially during the later speeches which could lead to teachers giving inaccurate grades to the students. The incorporation of PA into TA may allow teachers to be more relaxed during speaking tests as they know that they have the PA to support their own grading. Another possibility is that teachers can use PA as a part of the formal assessment procedures for speeches delivered in a group. As was suggested by some researchers (e.g. Fukazawa, 2007), in a regular English class it is very difficult to conduct speaking tests many times because assessing an individual student takes too much time. However, if we conduct a speaking test in groups it will save a lot of time. Luoma (2004) states that "peer evaluation is useful because it allows teachers to share some of the rating responsibility with their students, and it is especially useful in speaking assessment, which is time-consuming if rated by one person only" (p. 189).

Despite the potential benefits of PA there seems to be an obstacle which prevents it from being more widely used by teachers. This being that "many people believe that student-derived marks could not be used in formal grading procedures because they would not be accurate enough" (Stefani, 1994, pp. 69-70). Also, the "fears of teachers about the lack of reliability or validity of peer assessment may act to restrict its use and, thus, deprive many students of its learning benefits" (Falchikov & Goldfinch, 2000, p. 288). Therefore, more extensive analyses of reliability or validity of student-derived marks should be made "to determine the extent to which peer and self- assessments could be used in formal grading procedures" (Stefani, 1994, p. 70).

In analyses conducted for the above said purpose the level of agreement between TA and PA is usually sought to find if "there is a very high probability that student marks are the same as staff marks for a given assignment" (Boud, 1989, p. 20). However, studies have shown contradictory results. Some have found a high agreement (Fukazawa, 2007; Hughes & Large, 1993; Miller & Ng, 1996), some have found a good agreement only with certain conditions (Patri, 2002), and others have observed a low agreement (Kwan & Leung, 1996; Orsmond, Merry & Reiling, 1996, 1997). Therefore, it is necessary to further investigate the level of agreement between TA and PA. This paper intends to add our observations to this inconclusive area by reporting the results of a comparison between TA and PA of students' speech presentations conducted in an EFL context.

There are some crucial points to be considered for successful peer assessment. Establishing a well-defined set of criteria is important to help students grade accurately and teachers need to ensure students understand what each criterion means. It is also important to find if there are any difficult areas for students to assess. Teachers have to be aware of these areas so that they can spend more time addressing them when they explain the criteria to the students, and also so that they are cautious about using PA in these areas for any formal grading. Luoma (2004) has noted that linguistic criteria may not be suitable for PA "because students are not as adept at language analysis as teachers" (p. 189). Thus we decided to explore areas of difficulty for learner assessment. We were also interested in finding what kind of opinions the students would have toward PA after they had actually practiced them. The following research questions were therefore investigated:

1. Is peer marking comparable to teacher marking when our method is used?
2. Are there any difficult areas for the students to make teacher-like assessments?
3. What kind of reactions do we get from the students regarding peer assessment?

Method

Participants

For this study, the participants were 88 first-year students at a national university in the Kanto area who belonged to five different departments (Education, Agriculture, Engineering, Science and Humanities). All participants were in the university's mandatory Integrated English Program (IEP). The IEP is a 30 hour general English course that focuses on developing four English language skills: reading, writing, speaking and listening. As policy stipulates, all students are placed into five levels based on scores from a general proficiency test, level one being beginners and level five being the most advanced. Each class has about thirty students and meets twice a week. The students who participated in our study were in level three of the IEP and all of them were taught by one of the authors of this paper.

Data collection procedures

Assessment criteria and format

The assessment criteria for this study were carefully established by the two teachers based on their experience, available information and the implications of the previous studies (Cheng & Warren, 2005; Council of Europe, 2001; Luoma, 2002). The assessment criteria list (see Appendix 1) utilizes five points of assessment and can be summarized as follows: Voice volume, Pronunciation, Eye contact, Fluency, Grammatical Accuracy and Content. Due to a technical difficulty, Voice volume was excluded from the statistical analysis. The scale for each of these criteria was measured on a scale from 1 to 5 (with 1 being the lowest and 5 the highest). In order to assure the students' clear understanding of the criteria, the assessment criteria list was written in both Japanese and English. The list was made to look as simple as possible so that the students could easily refer to the criteria while listening to a speech.

Explaining the criteria to the students

Each assessment criteria was explained by the teacher using a prepared checklist (see Appendix 1) with the aid of teacher demonstrations (see Appendix 2). Some researchers recommend showing sample videos and highlighting the elements of good and bad presentations as a way of outlining criteria (Freeman, 1995; Patri, 2002). This, however, has practical difficulties. Model videos can be difficult to make or find. It is not easy to receive positive replies from the students from the previous years about the use of their taped speeches as a model and suitable exemplars are not always available. Therefore, we created a simple way of explaining the criteria where instructor's demonstrations were provided together with verbal explanations. In the first session, a full and detailed explanation of the criteria was given and in the following sessions simpler explanations were given as the students got used to the criteria.

Training for assessment

In assessment there is a need for "rigorous training and standardization of markers in order to boost test reliability" (Weir, 1990, p.80). Weir also adds that "the purpose of standardization procedures is to bring examiners into line, so that candidates' marks are affected as little as possible by the particular examiner who assesses them" (p. 82). Patri (2002) observes that "if learners are put in a situation where they can access information regarding the quality and level of their own performance, or those of their peers, then they will be able to clarify their own understanding of the assessment criteria" (p. 111). Taking these observations into consideration we decided to give the students four training sessions before the final presentation. We expected that the students would familiarize themselves with the grading process through the training. The four preparatory drills also served as important opportunities for the students to learn from peer feedback and to practice speaking in English on different topics.

Speeches and assessment

Each student made five speech presentations in total. All speech topics were chosen from *Interchange student book 3* (Richards, 2005), which is the text book for IEP level three classes. In each of the first four training sessions the students delivered a one to two minute speech to the other members of their group followed by one minute of evaluation time. The students made assessments by completing an assessment form (see Appendix 3) and these assessments were shared with the group before being collected by the teacher. In the final session, a two to three minute speech was delivered to the whole class, followed by one minute of evaluation time. This time each speech was evaluated by both the teacher and the students. The students were told that their assessment would not be read by the other students and would be submitted directly to the teacher who would later check if they had assessed their peers appropriately.

Videotaping

It is considered that "even with careful training, a single scorer is unlikely to be as reliable as one would wish" (Hughes, 1989, p. 114). In order to obtain a reliable benchmark for comparison with PA we needed more than one teacher to assess the students' performance. Therefore, the final (fifth) speech was videotaped by the class teacher (one of the authors of this paper) so that the other teacher (the other author of this paper) who was not in the classroom and did not directly observe the speeches, could evaluate all the speeches in the absence of the class teacher without knowing the scores given by her.

Students

After completing the peer assessment of the final speech, the students were asked to fill out a questionnaire regarding peer assessment (see Appendix 4).

Data Analysis

First, the average scores of markings of the two teachers were obtained to be used as a reliable benchmark with which we would compare the grades awarded by the students to their peers. We did this because reliability of a rater's judgment is believed to "be enhanced with multiple staff assessors" (Freeman, 1995, p. 291). Then, all of the TA average scores and PA scores except for Voice Volume were entered into an Excel spreadsheet. The resulting data was analyzed using SPSS to find the degree of agreement between TA and PA. In order to compare ranges of markings, the standard deviations of TA and those of PA were also calculated. The students' responses to the questionnaire were also entered into an Excel spreadsheet for analysis.

Results

1. Is peer marking comparable to the teacher marking when our method is used?

In order to see if there is good agreement between the TA and PA, Pearson's correlation tests were used. Table 1 shows the result of Pearson's tests for the mean marks awarded to each student by the teachers and by the students. Other than Grammatical Accuracy (r = 0.31), PA for each criteria consistently showed very strong correlations with TA. The overall correlation coefficient was as high as (r = 0.82) which suggests that students can be reliable assessors and that PA can supplement TA to some extent. The result we obtained is close to the result (r = 0.85) observed in one of the two groups in Patri's (2002) study where a sample video was shown to both of the groups to clearly establish criteria set by the researcher.

Table 1. Correlations between the Teacher marks and Peer marks

	Correlation coefficient (r)
Pronunciation	0.70**
Eye Contact	0.84**
Fluency	0.75**
Grammatical Accuracy	0.31**
Content	0.81**
Overall Total	0.82**

**Correlation is significant at the 0.01 level (2-tailed).

The number of the students: 88

Table 2 shows the means and standard deviations of marks awarded for each of the assessment criteria by TA and PA. The results show the students' tendency to give higher scores than their instructors, which is a phenomenon noted in previous studies (Freeman, 1995). The table also shows that standard deviations of the students were consistently smaller than those of the teachers, reflecting the tendency of students' using a narrower range of marks than their instructors (Cheng & Warren, 2005; Freeman, 1995; Hughes & Large, 1993). Cheng & Warren (2005) have noted that this "is usually ascribed to the reluctance on the part of students to mark their peers up or down" (p. 105).

Table 2. Mean and Standard Deviation for Teacher marks and Peer marks by criterion

	Teacher marks		Peer marks	
	M	SD	M	SD
Pronunciation	3.43	0.53	4.03	0.32
Eye Contact	3.26	0.84	3.7	0.48
Fluency	3.72	0.59	3.99	0.44
Grammatical Accuracy	3.51	0.63	4.33	0.2
Content	4.13	0.74	4.23	0.41
Overall total	18.05	2.38	20.29	1.48

The number of the students: 88

2. Are there any difficult areas for the students to make teacher-like assessment?

Varied degrees of correlations were observed for individual assessment criteria as shown in Table 1. Grammatical Accuracy was found to be very weakly correlated (r = 0.31), indicating that this is a difficult area for the students to make a correct assessment. Pronunciation, which is another linguistic area, was found to be slightly less correlated (r = 0.70). On the other hand, non-linguistic areas such as Eye contact (r = 0.84) and Content (r = 0.81) were strongly correlated. These results are consistent with aforementioned findings (Luoma, 2002).

3. What kind of reactions did we get from the students regarding peer assessment?

A 6-point scale (Level 1 = strongly agree, Level 6 = strongly disagree) was used on the questionnaire. The students reported a high level of confidence in their understanding of the meaning of each criterion (Level 1: 21%, Level 2: 45% and Level 3: 19%). This high level of confidence seems to back up the strong correlation (r = .82) between TA and PA and confirms the importance of clear marking criteria pointed out by Orsmond, et al., (2000, 2002) and Patri (2002). A majority of the students perceived the PA as useful (Level 1: 25%, Level 2: 32% and Level 3: 28%). This compares well with the result obtained in Orsmond's (2000) study where 80% of the participants reported that self/peer assessment was helpful.

More than half of the students reported a tendency to be lenient about scoring when a speech was done in a group (Level 1: 12%, Level 2: 23% and Level 3: 32%), and a majority of the students found it easier to make an assessment when a speech was delivered to the whole class than to a group (Level 1: 26%, Level 2: 32% and Level 3: 20%). During the training sessions they showed their assessment sheet to the group members before it was collected by the teacher. However, when a speech was delivered to the whole class, the sheet was not shown to others. This seems to be a cause of the difference in easiness in assessing their peers. The responses for these two questions imply that peer assessment is more reliable when students do not show their assessment to their peers and that it is not safe enough to incorporate formative PA with feedback from their peers into the final score.

Of the 88 students, 72 answered an open-ended question: *We have done five speeches so far - what did you think about the peer assessment?* Of them, 55 students (76%) gave positive comments on PA such as "We learned a lot from each other" and "I found both my strong points and weak points", and 11 students (15%) gave negative comments. Of the 11 negative comments, two students wrote "I felt nervous" and five students wrote "It was difficult to make assessments."

Conclusion

Our study found that the responsibility for assessing students' speech presentations may be shared by the teacher and the students. A strong correlation (r = .82) was observed between TA and PA in the final presentation after four practices. This implies that it is viable to incorporate PA into the formal grading procedures when our training method is applied. Whether we can use PA conducted in a small group as a part of the final score was not examined in this study. This is a question worth investigating. Our study also found that the equivalency between TA and PA varied in strength, and that agreement was not great in criteria which involved linguistic rules. Thus, we may assume that the teacher's marks should hold precedence in these areas. A questionnaire conducted after the final PA revealed that most of the students perceived benefits of PA indicating that PA can be a positive educational tool.

While we obtained some interesting results, the small range of the proficiency of the participants of this study prevents the generalization of the findings. In order to see if our method can be applicable to a wider range of students further research needs to be conducted. Another weakness of this study lies in our questionnaire. In using the questionnaire we prepared, we were not able to clarify how the students felt about being assessed by their peers and how the students would feel about having these assessments used in their final grades. These points need to be explored in a future study. We hope that our findings will help promote the use of peer assessment in English language education classes.

Acknowledgements

We would like to express our appreciation to Yo Innami and Makoto Fukazawa for their support and invaluable suggestions.

References

Boud, D. (1989). The role of self-assessment in student grading. *Assessment and Evaluation in Higher Education, 14*(1), 20-30

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

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22*, 93-121.

Clifford, V. A. (1999). The development of autonomous learners in a university setting. *Higher Education Research & Development. 18*(1), 115-127

Council of Europe (2001). *Common European Framework of Reference for Languages; Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20*(3), 289-299.

Fukazawa, M. (2007). *Validity of Peer Assessment of Speaking Performance: A case of Japanese High School Students.* Unpublished master's thesis, University of Tsukuba, Japan

Hughes, A. (1989). *Testing for language teachers.* UK: Cambridge University Press

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

Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment & Evaluation in Higher Education, 21*(3), 205-215.

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

Miller, L. and 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.

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.

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment; tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education*, *22*, 357-369.

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, *25*, 23-38.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, *27*, 309-323.

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

Richards, J.C., Hull, J. & Proctor, S. (2005). *Interchange student book 3.* Cambridge: Cambridge University Press.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*, 69-75

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

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

Rieko Okuda is a part-time English instructor at Hiroshima University and Hiroshima Shudo University.

Rika Otsu is currently pursuing a graduate degree in TESOL at California State University, Fullerton.

Appendix 1: *Assessment Criteria*

Category	**Level**	Description
Volume of Voice	1	difficult to hear
	3	sometimes difficult to hear
	5	easy to hear
Pronunciation	1	not natural
	3	sometimes not natural but does not affect the speech delivery
	5	natural and appropriate
Eye Contact	1	does not look at listeners
	3	sometimes looks at listeners but not everyone
	5	always looks at listeners
Fluency	1	too many pauses
	3	some unnecessary pauses or hesitations
	5	smooth without hesitation
Grammatical Accuracy	1	too many grammar or usage mistakes
	3	A few grammar and usage mistakes
	5	almost correct grammar and language use
Content	1	Subject content is not clear and lacks sufficient information.
	3	Content is clear but needs more information.
	5	Clear content and sufficient information.

Appendix 2: *Instructions given to the students*

Pronunciation:

“1” point will be given to a speech with a flat intonation with katakana English sounds where every consonant sound is followed by a vowel sound, whereas, “5” points will be given to a speech with natural English sounds. “3” points will be given to a speech in between Katakana English and natural English sounds.

Eye Contact:

If a speaker keeps looking downward and does not try to keep eye contact with listeners, “1” point will be given. If a speaker constantly remains in eye contact with the listeners throughout the speech, “5” points will be given. “3” points will be given to a speaker who sometimes tries to keep eye contact but sometimes looks downward to check the script.

Fluency:

A speech with many unnecessary pauses or hesitations will get “1” point. A speech at a natural speed without unnecessary pauses or hesitations will get “5” points. “3” points will be given to a speaker who sometimes makes unnecessary pauses unintentionally.

Grammatical Accuracy:

A speech that contains many grammatical errors causing difficulties for listeners to understand the speech (e.g., Me, friend with, Disneyland, went to go, summer, before, last year) will receive “1” point. “5” points will be given if a speech is, for the most part, grammatically correct. (for example, I went to Disneyland with friend last summer.). “3” points will be given to a speech if it has some minor grammar errors but can be easily understood (for example, I went to go to Disneyland with my friend last year’s summer.).

Content:

“1” point will be given if a speech is difficult to follow and does not have enough information. An example of a speech with insufficient content would be: “My girl friend and I drank beer. She got angry. That’s why, I had a bad day.” If a speech is clear and concise and has enough amount of information, “5” points will be given. For example, “I had a bad day today, because I had a terrible argument with my girl friend. After school, we went to a bar and started to drink. I drank a little too much and started to complain about our friend, Takashi. I said …… My girl friend got angry because she did not agree …….” “3” points will be given if a speech is not clear or descriptive enough to tell the whole story but is still understandable and predictable. For example, “I had a bad day. I had an argument with my girlfriend. We drank beer and I complained about our friend. So, she got angry.”

Note: Appendix 3: *Assessment Sheet*, and Appendix 4 *Speech Assessment Questionnaire* are available online at <jalt-publications.org/tlt/resources/2010/04a.pdf>.