EFL classroom peer assessment: Training effects on rating and commenting

Hidetoshi Saito

Language Testing 2008 25: 553
DOI: 10.1177/0265532208094276

Language Testing 2008 25 (4) 553–581

EFL classroom peer assessment: Training effects on rating and commenting

Hidetoshi Saito *Ibaraki University, Japan*

This study examined the effects of training on peer assessment and comments provided regarding oral presentations in EFL (English as a Foreign Language) classrooms. In Study 1, both the treatment and control groups received instruction on skill aspects, but only the treatment group was given an additional 40-minute training on how to rate performances. The results of the correlation difference analyses show no significant differences between the treatment and control groups, but the three instructors are misfitting, presumably because the frame of reference is set by the majority of student data. In the second study, the treatment groups received long training. Again, there are no significant correlation differences between the treatment and control groups; however, all the instructors are not misfitting, which indicates that the frame of reference in the data improved in Study 2. Analyses of comments reveal that the treatment groups are superior in both quality and quantity of comments. Along with a meta-analytic summary, it is argued that peer assessment is a robust system in which instruction on skill aspects may suffice to achieve a certain level of correlation with the criterion variable (instructor), but training may enhance student comments and reduce misfitting raters.

Keywords: comments, effect sizes, peer assessment, rater training

I Introduction and literature review

The number of empirical research studies on peer involvement in classrooms has increased. Some researchers have claimed that working with peers in the classroom is a critical means of promoting learning. Such claims draw on evidence from second language (L2) acquisition research, mainstream education research, and both L2 writing and L1 writing research (e.g. DiPardo & Freedman, 1988;

Address for correspondence: Hidetoshi Saito, Ibaraki University, Department of English, College of Education, 2–1-1 Bunkyo, Mito 310–8512, Japan; email: cldwtr@mx.ibaraki.ac.jp

 DOI:10.1177/0265532208094276

Liu & Hansen, 2002; Long & Porter, 1985; Webb, 1982, for review). Peer involvement in assessment holds tremendous potential for learning as well. The literature from various fields offers general agreement on the following characteristics and potential benefits of peer assessment (including the so-called 360 degree assessment in industrial/ organizational psychology):

1. Although evidence drawn from language classrooms is limited, peer assessment has a fairly strong correlation with instructor ratings across different subject areas and with supervisor ratings across different organizational settings (e.g. Falchikov & Goldfinch, 2000; Fletcher & Baldry, 1999; Harris & Schaubroeck, 1988; Topping, 1998).
2. Peer assessment encourages reflective learning through observing others' performances and becoming aware of performance criteria (Falchikov, 1986; also claimed in different contexts by Topping, 1998; Tornow, 1993; Somervell 1993).
3. In general, peer assessment seems to generate positive reactions from students, although some students have concerns and worries, as demonstrated by the mixed findings in various content areas (e.g. Cheng & Warren, 1997; Haaga, 1993; Morahan-Martin, 1996; Saito & Fujita, 2004; Stefani, 1992).
4. Students develop a sense of shared responsibility (e.g. as claimed by Somervell, 1993).

The benefits which peer assessment may bring into a language classroom cannot be guaranteed unless students are capable of implementing the assessment. The same concerns about student capacity to discern peer performance and the need for training were raised in recent guidelines on peer review in L2 writing (Hu, 2005; Liu & Hansen, 2002). In fact, several research studies on peer writing response groups have investigated the capacity issue by looking at training effects and have found benefits of training for the revision process. In these studies, researchers compared trained and untrained groups by categorizing and counting the frequency of peer comments. The trained group produced more specific responses than the untrained group, and the trained students were more responsive to comments in revision relative to untrained students (Stanley, 1992). In particular, training has been found to lead to more changes of 'meaning' in revision (Berg, 1999). Similarly, in predominantly L1 English freshman composition classrooms, trained groups were found to spend more time on revision and to produce more relevant comments than those who were untrained (McGroarty & Zhu, 1997). The effects

of training on the improvement of the final product were found in Berg's study (1999) but were absent in a study by McGroarty and Zhu (1997). Although the advantages of training for peer writing responses appear to be confirmed, no study has required students' use of a rating scale of any kind. Thus, it is not clear whether training induces the same benefits for peer rating or whether it works for oral performance in the same way as it does for writing.

Rater training has played a critical role in assuring rater consistency and agreement on performance assessments (especially in high-stakes tests). Several L2 language testing researchers have examined training effects in relation to both expert and novice raters. These studies have offered evidence that supports the following: (1) trained raters are more reliable than untrained raters (Shohamy, Gordon & Kraemer, 1992; Weigle, 1994, 1998) and (2) training makes raters more self-consistent but does not dramatically alter severity (Lumley & McNamara, 1995; Weigle, 1998). However, research is inconclusive about what roles background variables (whether evaluators are lay or experienced) may play in relation to training. For example, Shohamy *et al*. (1992) found positive effects of training but no background effects in terms of reliability differences. In Weigle's studies (1994, 1998), novice raters were more strict and inconsistent than experienced raters, but training reduced extremism in novice ratings.

In other L2 studies, mainly those on writing, sources of difference between experienced and novice raters have been identified. For example, Cumming (1990) suggested that a major difference between lay and experienced raters in the rating process lies in the range of criteria, strategies, and knowledge resources they could respectively exploit. Not surprisingly, however, experience does not always lead to a strong consensus among raters (Lumley, 2002). Schoonen, Vergeer and Eiting's (1997) study, although not focused on training, has drawn a complex picture of the role of expertise in rating, partly because of the research design utilized. Schoonen *et al*. (1997) compared the reliabilities of lay raters with those of experienced raters on three different writing tasks using structural equation modeling. Although the results suggested that both groups differentiated writing products in a similar way, the researchers concluded that several factors, such as differences in writing tasks and rated aspects, may affect the differences in severity between lay and experienced raters.

These studies provide insight into the complex interaction between rater training and experience, but the novices in these studies were not

L2 learners. L2 learners differ from the lay raters in the studies above in at least three respects. First, lay raters in those studies were native speakers; whereas, L2 learners in peer assessment studies were not native speakers of the target language. It is likely that their developing language ability affected the peer ratings in various ways. Second, raters in those studies were anonymous; whereas, learners in peer assessment studies often know who they are rating or are at least aware that they are rating their peers. This awareness is likely to generate leniency and uneasiness, which may not be a factor in the lay/experienced rater comparison studies. Third, the products and performances rated in peer assessment studies have been strongly associated with actual classroom practices; whereas, the settings of lay/experienced rater comparison studies are independent of classroom learning. The lack of understanding of how these factors interact with peer assessment further justifies investigating the effects of training on L2 peer rating.

Studies on L2 peer assessment have shown some evidence in support of its use in classrooms. Some studies have shown that peer assessment is correlated with instructor assessment (Jafarpur, 1991; Patri, 2002; Saito & Fujita, 2004; Yamashiro, 1999), while others have found correlations to vary depending on tasks and classes (Cheng & Warren, 1999). Peer feedback increases the correlation between peer ratings and instructor ratings (Patri, 2002), although marking in general may be inflated (Cheng & Warren, 1999; Patri, 2002; Saito & Fujita, 2004). Regarding user acceptance, students have shown both generally positive (Rothschild & Klingenberg, 1990; Saito & Fujita, 2004) and mixed attitudes (Cheng & Warren, 1997) toward the use of peer assessment; however, such differences in attitude do not appear to be related to the feedback received (Saito & Fujita, 2004). Unfortunately, none of these studies have investigated the effects of training on peer rating. Moreover, it is clear that these findings seem somewhat contradictory and, consequently, do not give solid advice to practitioners. Due to such limited and diverse findings, a quantitative summary of available studies invoking prior effect sizes may situate the present study in the proper context of related research, as recommended by researchers such as Thompson (2002). Effect size indicates the degree of association between variables and is considered to be appropriate for describing the actual significance of a study. The remainder of this section offers a quantitative summary based on effect sizes (i.e. the strength of the relationship between teacher and peer ratings) retrieved from four empirical studies. To explore the relevant literature, the ERIC database (1966–2003) was searched using ‘peer’, ‘language’, and

'assessment' or 'rating' as keywords. Because the present study intends to identify a status quo based mainly on accessible studies, this search was limited to journal articles and book chapters; however, by excluding what is called 'fugitive literature', this search might reflect a 'publication bias' by only including results that are published and widely available (e.g. Rosenthal, 1991). Thus, readers are cautioned that this quantitative summary is by no means complete and is biased in favor of published studies. A manual search of book chapters and reference works was also implemented.

Four quantitative studies (out of seven located) met the following criteria:[1] The study (1) describes necessary information for summarizing, including sample size, settings, and data collection procedures; (2) provides observational data by both instructors and L2 learners ratings on peer products or performance using a rating scale of some sort; and (3) shows correlation(s) or other measures that can be transformed to an adjusted r^2. To summarize the effect sizes of the studies, adjusted r^2s were retrieved or calculated. Since each study usually generated more than a single r, averaging (with Fisher's z transformation) was done so that each study contributed only one r^2 to the present summary (see Rosenthal, 1991). The results of the mean adjusted r^2 and 95% confidence intervals (CI) appear in Table 1.

To test the heterogeneity of the adjusted r^2 retrieved from the four studies, the chi-square was calculated. The x^2 value of 1.15 ($df = 3$) was not significant at .05, indicating that the effect sizes were not heterogeneous. In other words, the effect sizes of the studies summarized here collectively show that the strength of the relationship between L2 peer and instructor ratings is fairly stable across studies.

As informative as the studies reviewed above may be, a number of questions still remain unanswered regarding the role of training in peer assessment in L2 classrooms and the effect of training on L2 students' ratings and comments, particularly for oral performances. The following hypotheses were formulated based on the previous studies:

1. Trained groups' peer ratings of presentation quality correlate with those of the instructors much more highly than those of untrained groups.
2. Trained groups make qualitatively better and quantitatively more comments on peers' performances compared to untrained groups.

---

[1]Three studies (Devenney, 1989; Jafarpur, 1991; Rothschild & Klingenberg, 1990) were not included because they did not meet one or more of the above criteria. Yamashiro (1999), an unpublished yet award-winning conference paper, is included here because it was presented at a major conference and met all the criteria.

Table 1 Results of quantitative summary of previous studies

Study	Averaged sample size[1]	Number of instructors	Averaged adjusted r^2	Lower 95 CI	Upper 95 CI	Research questions	Skills assessed	Rater training	Context	Main results
Cheng & Warren (1999)[2]	17	3	.265	.000	.535	Identifying characteristics	All class seminar, small group oral presentation, written report	Yes	University students in Hong Kong	Correlations varied depending on task types and classes
Patri (2002)	27	1	.500	.197	.668	Examining effects of feedback on rating	Small group presentation	Yes	University students in Hong Kong	Peer feedback helps improve rating and improves correlations with instructor
Saito & Fujita (2004)	50	2	.513	.303	.643	Identifying characteristics and effects of attitude on rating	Essay writing	Yes	University students in Japan	Item hierarchies between instructor and peers differed. No attitude effect on rating was found.
Yamashiro (1999)	170	7	.514	.411	.593	Validating a rating scale	Public speaking	Yes[3]	University students in Japan	The validity of a rating scale based on peer, instructor, and self-ratings was supported
Pooled	264	13	.501	.403	.587					

Notes: CIs (confidence intervals) were calculated using an SPSS script by Smithson (n.d.).
[1] The number of learners rated. [2] This study is believed to be identical with Cheng and Warren (2005), so the latter study was not included.
[3] Not documented in the paper, but it is likely that students followed the training procedure described in Yamashiro and Johnson (1997).

In the following, two studies are described. The first study, with a short training time, examined the first hypothesis, while the second study tested both hypotheses. The results of the first study are described briefly due to space limitations. The second study is a replication of the first study, but a longer training time is implemented.

II Study 1

1 Method

a Participants: After eliminating absentees, 74 Japanese university freshmen, all majoring in economics, who were registered in three sections of a speaking-listening course, participated in the current study. Following the national curriculum, all had completed six years of English studies (approximately 750 hours) in middle school. These three sections were chosen because, based on an in-house placement test, the proficiency levels of the students were considered relatively high and similar to each other. Two sections were considered as the second best groups, and one section was the third best. In the course, all three sections used the same textbook and were given the same mid-term and final exams. Twenty-two students were randomly subjected to SST (a subset of ACTFL OPI), and all fell into a mid-novice to high-novice range. None of the participants reported previous experience in assessing peers' English presentations in the background questionnaire. Three instructors, all Japanese teachers of English with greater than five years of experience teaching at the university level and experience living abroad for more than two years, also participated as raters in the study, including the instructor who taught all three sections of the speaking-listening course.

b Procedure: As part of the course requirement, students were assigned an oral presentation on a topic of their own interest, such as travel, hobbies, or club activities. They were told that this presentation and the presentation draft accounted for 15% of their grade. The students were also told that they needed to speak for at least 5 minutes and were required to use visual aids. Both the treatment and control groups received a series of instructional inputs on 12 skill aspects of presentation (hereafter, called instruction) directly associated with the assessment items, but only the treatment group had a rater training session after the fifth session. Each instruction session (for the first five sessions) began with the instructor demonstrating good and poor examples of each skill aspect. For example, in the first session, the instructor demonstrated strong and weak types of gestures and postures for an oral presentation, and the students

practiced the gestures and postures through paired activities (see Appendix 1). This instruction on performance aspects continued until the fifth session. After the last session, each of the three sections was randomly divided into treatment ($n = 37$) and control groups ($n = 37$). The treatment group moved to a different room and had a rater training session for approximately 40 minutes, after which video recordings of three presentations by former students were viewed, while the control group stayed in the classroom and was engaged in an irrelevant writing assignment.

Rating practice consisted of the following steps: First, the instructor explained all of the items in detail. Students viewed and rated three videotaped presentations of former students, each of which represented superior, adequate and minimal levels of performance, or required work on more than two skill items. The students compared their ratings with those of others in the group and also reported, by raising their hands, what rating they had given to each item in order to compare their ratings with that of the instructor. The instructor then explained why a certain rating was more appropriate for an item, pointing out and discouraging obvious over- and under-rating. This comparison and check process was repeated after viewing each videotaped presentation. Due to time constraints, this training provided a selective benchmark performance in order to illustrate ratings for each item. Although such quick and short training was not ideal, it was assumed to represent the time allotment that many language teachers can afford in their actual courses. Instructors also received the same rater training for about 40 minutes, which was given by the instructor who also taught the students.

In the sixth to eighth sessions (see Appendix 1), students rated and commented on all classmates' performances in their own class using PASS (Peer Assessment Support System), an Internet-based program. In the ninth session, the students received feedback from their peers. Due to logistical reasons, instructors rated videotaped performances rather than live presentations. It was assumed that this method of presentation had negligible influence on the data; however, this mode difference may have affected the results, as discussed by McNamara and Lumley (1997). Instructors were not allowed to rewind the tape except for a few instances in which they felt that they had not concentrated enough to judge the performance.

c Instruments: The original form of the oral presentation skills assessment used for this study was developed by Yamashiro (1999) and Yamashiro and Johnson (1997). Oral presentation skills constitute a set of discrete behavioral aspects of task performance in language

learning. Students learn these skill aspects in classrooms through practice and observation. The instrument was originally composed of three aspects: verbal delivery, non-verbal delivery, and organization/purpose, each of which contained four skill aspect items. Yamashiro (1999) put this instrument through a multitrait-multimethod analysis, and the data supported the validity of the scale, generating an average effect size (adjusted r^2 of .51) (Table 1). This scale was slightly modified for the present study in two respects. First, item Purpose was replaced with Visual aids because Purpose seemed to be vague and impressionistic rather than observational, and the use of visual aids was emphasized in instruction. Second, the number of steps (rating category) was reduced from five to four because the researcher wanted to avoid the middle category that often attracts most responses and also believed that distinguishing among five levels for 12 items would be too daunting a task for the students. Prior to Study 1, the pilot data for this modified version were analyzed using Rasch analysis, with all items considered at once. Two items, Gestures and Visual aids, were found to be misfitting. Thus, separate runs for visual aspect and verbal aspect were executed, with the Rasch reliabilities for both aspects being .98. Although Diction was found to be slightly misfitting in this separate run, these two aspects are used for analysis in the following two studies (see Appendix 1). The verbal aspect constitutes skills that make the delivery of messages with words possible; whereas, the visual aspect refers to skills that allow the delivery of non-verbal messages by the speaker.

d Analysis: Rasch analysis is a test analysis method that carries several advantages over a classical test analysis, one of which is the separate calibration of measurement facets. In the present study, Rasch analyses were used to calibrate three measurement facets – item difficulty, rater severity, and presentation quality – separately for each verbal and visual skill aspect.[2] Rasch analysis requires unidimensionality of the data, so the visual and verbal dimensions were analyzed separately. Another advantage of Rasch analysis is the use of fit statistics to flag any aberrant responses in the data. These 'misfitting' responses are identified through criteria such as those proposed by Wright and Linacre (1994).

A series of one-factor linear regression analyses were used to predict instructor rating (dependent variable) from peer rating (independent variable). This analysis was done for presentation quality

[2]When analyzing the data, all of the data for each aspect were entered first. While anchoring rater severity and step measures generated in the first run, separate runs for each rater group were done to calibrate quality and item measures. This was necessary to ensure the connectivity of the data.

measures for both verbal and visual aspects. Note that all of the student variables abide by the normality requirement, but the instructor ratings for the two aspects were not found to be normally distributed. Thus, instructor ratings in the verbal aspect were normalized through reflection and a square root transformation, while the instructor ratings in the visual aspect were normalized through a square root transformation alone. Subsequent residual plots indicated that all dependent and independent variables had a linear relationship. Finally, to test the first hypothesis, the magnitude of correlations between treatment-instructor and control-instructor were compared using dependent correlation difference analyses in the ZumaStat program. Correlations were dependent because both the control and treatment groups rated the same classmates.

2 Results and discussion

The initial calibration of both the visual and verbal aspects produced high Rasch reliabilities across the three facets (.94–1.00). A closer look at each facet reveals that two peer raters in the visual skills do not fit the Rasch model, while four peer raters (three of whom were in the control group) and the three instructors do not fit the Rasch model in the verbal skills aspect. Among individual items, Diction is found to be slightly above the misfit criteria (infit/outfit = 1.4, $z = 9$). Although the overall reliabilities are high, it is surprising that all three instructors were misfitting in verbal skills assessment (infit/outfit = 1.7–2.0, $z = 9$). One possible explanation for this could be that the frame of reference set by peer learners, comprising 88.4% of the data, differ from that of instructors. If instructors and peers possess different frames of reference, the pattern set by the minority group (instructors) would not fit into the pattern of the majority (peers). Such differences in the frame of reference seem to be more apparent in verbal skills than visual skills. The main purpose of this analysis is to describe what is observed rather than to construct a high-stakes test, so none of the misfitting items or raters are eliminated in the subsequent analyses.

The results of the one-factor linear regression analyses were used to test the predictability of instructor ratings (dependent variable) from peer ratings (independent variable) for the presentation quality measures. All regression analyses generated statistically significant *b* parameters (the slope of the regression line), along with medium to large effect sizes of adjusted r^2 for both groups, .24 (verbal aspect in control), .32 (verbal in treatment), .71 (visual in control), and .69 (visual in treatment).

The magnitude of correlations between treatment-instructor and control-instructor were compared using dependent correlation difference analyses to test the first hypothesis. The results are $t = 1.10$ ($df = 71$, $p > .05$) for the verbal aspect and $t = .43$ ($df = 71$, $p > .05$) for the visual aspect. Both comparisons fail to reach statistical significance, and the two groups show no difference.

The results of Study 1 indicate that, since both groups are equally good predictors of instructor rating, there are no significant correlation differences between treatment and control. It is also noteworthy that the instructor ratings are misfitting, possibly because the frame of reference in the data set by the majority peer groups differs from that of the minority instructor group. The lack of differences between the treatment and control groups, as well as the misfitting problems, have led to the question as to whether longer training may resolve these problems.

III Study 2

1 Method

a Participants: After eliminating absentees, a total of 81 Japanese freshmen (ages 17–19) participated in Study 2. Based on the self-reported background questionnaire, no students had previous experience in assessing peers in language classrooms. The students were registered in three different sections of the speaking-listening course at the same Japanese private university as in Study 1. The treatment of Study 2 spread over 5 sessions, making it impossible to randomly divide each section into treatment and control groups. Thus, three intact groups participated in the study. The two treatment groups belonged to the Economics (hereafter Treatment 1) and Social Welfare (hereafter Treatment 2) Departments, and the control group students were Applied Psychology and Communication Studies (hereafter Control) majors. All three groups belonged to higher level sections within each department. Student assignment to a section was based on the institutional placement test administered before the semester had begun. In this program, section assignment was nested within each department for administrative reasons. The general proficiency of the students, as measured by the placement test, shows a significantly different $F\ (2, 78) = 4.89$, $p < .05$. The results of multiple comparison Scheffé tests suggest that the difference lies between Treatment 1 and Treatment 2. Note, however, that the actual raw means do not greatly differ (Treatment 1 $M = 45.30$, $SD = 2.57$; Treatment 2 $M = 43.12$, $SD = 1.34$; Control $M = 44.62$, $SD = 3.28$). In the course, all students used the same textbook and were given

the same mid-term and final exams. Four instructors, all Japanese teachers of English, with greater than five years' experience teaching at the university level and greater than two years' experience living abroad, also participated as raters in the study, including the instructor who taught all three sections of the speaking-listening course.

b Procedure: As in Study 1, both the treatment and control groups received a series of instructional inputs on 12 skill aspects of presentation. This time, however, the treatment groups had a rater training session after each instruction (see Appendix 1). Following the instruction on skill aspects, the treatment groups received approximately 40 minutes of rating practice. For example, in the first session, two skill aspects, gestures and posture, were practiced during instruction and were, thus, the focus of that day's rater training. The instructor explained each item in detail, and the students observed video clips of two former students' performances that were representative of two benchmark responses in Gestures and Posture. After the first video was viewed, individual students rated the two skill aspects using the rating scale. They then formed groups of three or four and compared and discussed their ratings by asking each other 'Why did you give a particular score?' Students were then asked to raise their hands to indicate what score they had given, and the instructor explained why a certain rating would be more appropriate based on the guideline. The instructor also pointed out and discouraged obvious over- and under-rating. This comparison and check process was repeated after viewing each videotaped presentation. The total training time amounted to about 200 minutes over five sessions, with 10 former student presentations viewed. The control group was engaged in textbook-based activities on speaking, listening, or grammar. Instructors also received approximately 90 minutes of training using the 10 video clips and benchmark responses used to train the students.

As in Study 1, students rated and commented on all other classmates' performances in the same class in the sixth to eighth sessions, and in the ninth session, students received feedback from their peers (see Appendix 1). Instructors again rated videotaped performances rather than live presentations. Instructors were not allowed to rewind the tape unless necessary.

c Instrument: The rating scale used was the same as the one in Study 1.

d Analysis: As in Study 1, Rasch analyses and a series of one-factor linear regression analyses were used. All of the variables conformed

to the normality requirement, and subsequent residual plots indicated that all dependent and independent variables had a linear relationship. Correlation difference analyses were used to test the first hypothesis. Difference tests for independent correlations were run because the ratings were done within each intact group, that is, they were independent of each other.

To examine the second hypothesis, comments given by students in the treatment and control groups were compared in three analyses: relevance judgment, overall frequency of cited items, and the frequency of each cited item. Due to technological failures, however, all comments of Treatment 2 and some portions of other groups' comments were inadvertently lost, leaving 593 comments (84.4% of all comments) made by respondents in Treatment 1 ($n = 26$) and 486 comments (74.7% of all comments) by respondents in the control group ($n = 27$). Since each peer rater provided comments for more than one peer performer, simple summing of the frequency counts was not appropriate due to the dependence in observation (Saito, 1999). Thus, all three analyses involve frequency or rating per comment averaged within a single peer rater. In relevance judgment analysis (following McGroarty & Zhu, 1996), two trained graduate students rated all the comments in terms of relevance to the performance, with the rating scale being 3 (relevant and specific), 2 (relevant but not specific), or 1 (irrelevant). The inter-rater agreement rate of relevance judgments was 92.5% (intraclass correlation of .71). All discrepancies were subsequently resolved through discussion. Relevance scores were averaged within a single peer rater, so each rater has a single mean relevance score.

In the second analysis, the overall frequency with which an item was cited was calculated, which is the number of skill aspects cited per comment averaged within each rater. A comment given to one performer may capture more than one category of performance. For example, the following comment made by one peer rater was categorized into Gestures, Body, and Pace: 'We could enjoy the talk because the speaker's gestures were very effective. The content was very interesting. It would have been much better had the speaker paid more attention to pace' (*T3002*). This comment describes three aspects, so the rater received a score of three for this one comment. The number of aspects cited was summed and averaged by the number of comments made by the single rater, so again, each peer rater had a single average frequency-of-citation score. Two graduate students independently categorized all of the comments into 12 skill aspects and proposed two additional categories: Loudness and

Pronunciation. After reclassifying the data with these two categories added, an agreement rate of 91.3% was reached. Subsequent discussions resolved all discrepancies. A MANOVA (multivariate analysis of variance) was used for group comparison of the relevance judgment scores and overall frequency of citation scores.

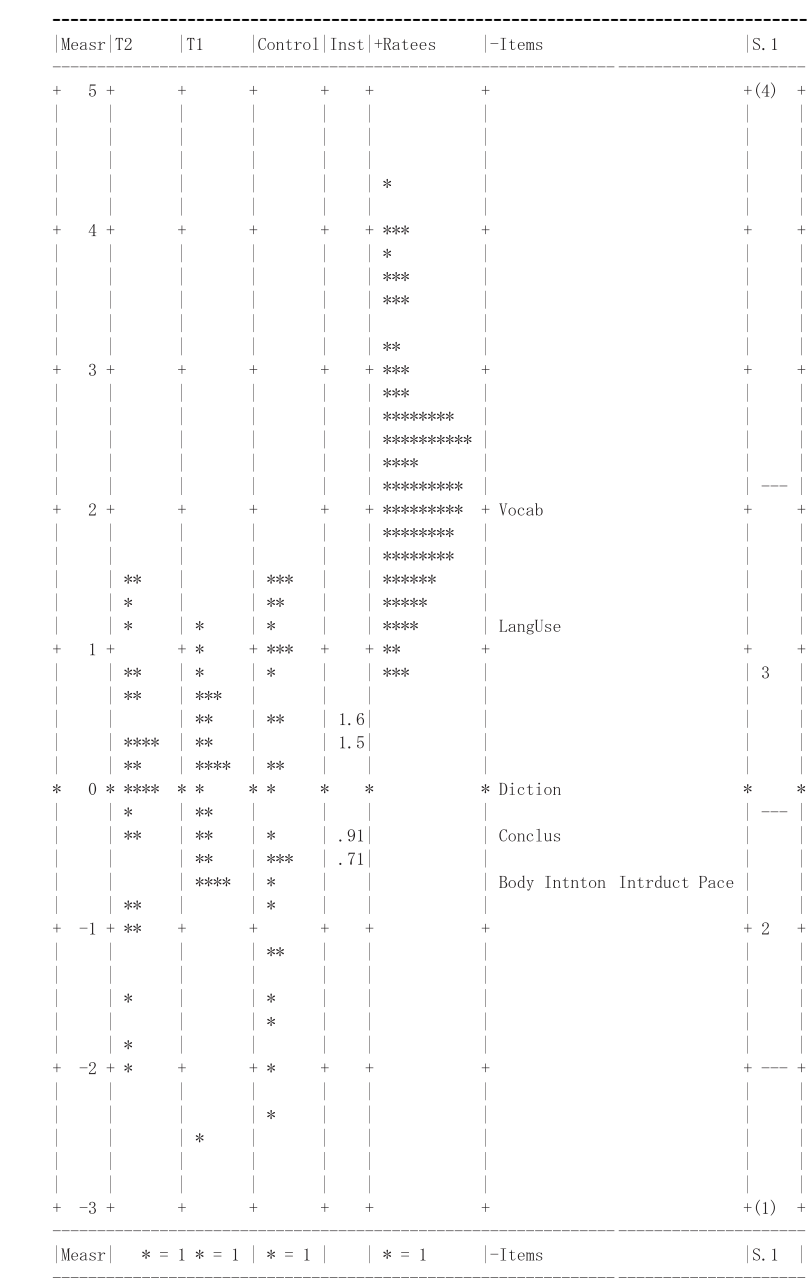
The third analysis of comments involved comparisons of the frequency of citation according to each skill aspect. Again, the frequency of skill aspect citations per comment was averaged within a single rater, so each rater had an average score for each skill aspect. There were many empty and small cells; that is, peer raters did not refer to all skill aspects when commenting. This factor naturally resulted in a lack of normal distributions of the mean scores. Thus, to test the difference between the groups, a non-parametric test, Mann-Whitney U, was used.

2 Results and discussion

Table 2 reveals that the Rasch reliabilities are fairly high across facets and aspects except for the presentation quality measures in the visual aspect. Reasons for this difference can easily be inferred from Figures 1 and 2, where items and raters are relatively more on target in the verbal aspect compared to the visual aspect. In both cases, however, presentation quality measures are much higher than raters and items. This result can be interpreted as the overall leniency of the raters or as the perception of most students having reached a certain level of achievement, since this was a classroom assessment.

Concerning fit statistics, all items except for Vocabulary were fitted to the Rasch model. As is the case with Diction in Study 1, the values of the fit statistics for Vocabulary were a little above the border when applying the criteria of Wright and Linacre (1994). No instructors were misfitting in Study 2, despite the fact that peer data again comprised 87.4% of the data. Only two peer raters were (one each from treatment and control) misfitting. No misfitting items or raters were eliminated in the subsequent analyses because this analysis was used mainly for descriptive purposes.

Table 3 shows the descriptive statistics for the presentation quality and rater severity measures. Both statistics indicate that instructors are more severe in their ratings than other groups in all respects except for the visual aspect in Treatment 1. Among the three peer groups, Treatment 1 had the lowest means on presentation quality and the most severe raters; whereas, Control had the highest means on presentation quality and the most lenient raters. Leniency



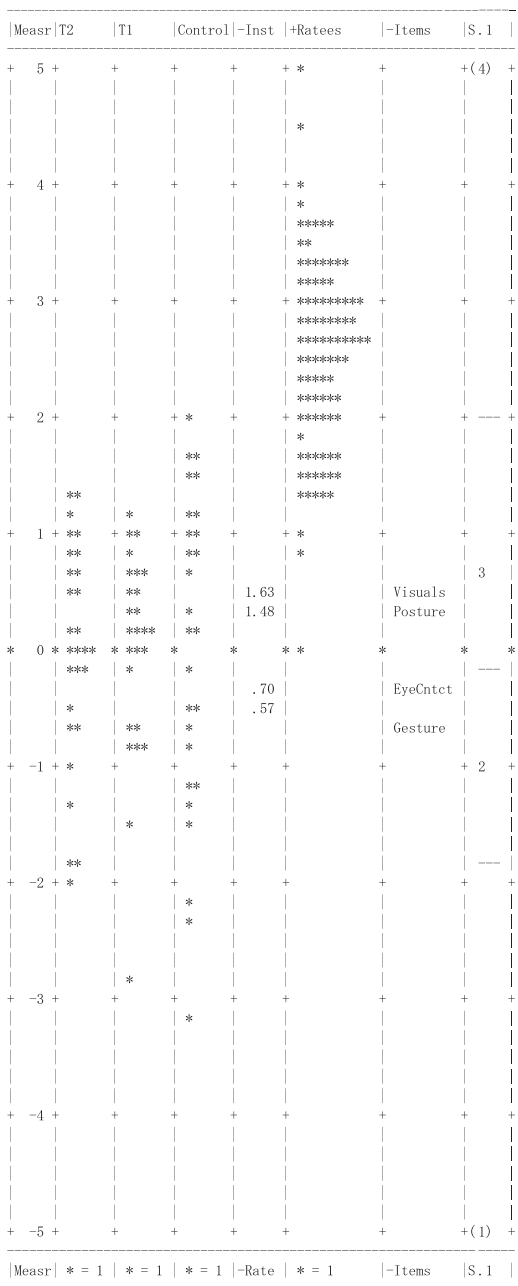
Notes: Measr = Rasch logit measures; T1, T2 = Rater severity measures of Treatment 1 and 2; Inst= Rater severity measures of Instructors; Ratees = Presentation quality measures; Items = item difficulty measures.

Figure 1 Results of Rasch analysis of verbal aspect

of peer ratings, in comparison with teachers, concur with several previous studies on L2 (e.g. Patri, 2002; Saito & Fujita, 2004) and in other domains (e.g. Morahan-Martin, 1996; Stefani, 1992). Note also that instructors' standard deviations in presentation quality were wider than in the student groups, indicating wider discrimination in rating.

The results of the regression analyses in Table 4 indicate that all groups are fairly good predictors of instructor ratings of presentation quality. However, in both aspects, the control group appears to be a slightly better predictor than the treatment groups. Correlation difference tests for independent correlations were also run, in order to examine differences between the treatment and control groups for each aspect. The results suggest that there are no differences between the control and treatment groups in either aspect, $z = .82$, $p > .05$, $r = .09$ for verbal aspects of Control and Treatment 1, $z = .48$, $p > .05$, $r = .05$ for verbal aspects of Treatment 2 and Control, $z = 1.63$, $p > .05$, $r = .18$ for visual aspects of Treatment 1 and Control, and $z = .66$, $p > .05$, r = .07 for visual aspects of Treatment 2 and Control. The results suggest that Hypothesis 1 is not confirmed. It is suspected that performance differences among the intact groups might have affected the results, although there were no differences in placement test scores. To check this possibility, a MANOVA was run to compare the treatment and control groups on presentation quality rated by the instructor. The results, however, were not significant: Wilks' $\lambda = .93$ ($F(4, 154) = 1.254$), $p > .05$, $\eta^2 = .032$. Thus, the absence of a training effect does not seem to be due to a difference in performance.

Concerning the first comment analysis, the descriptive statistics of the relevance judgment were $M = 2.78$, $SD = .32$ for the control group and $M = 2.93$, $SD = .14$ for the treatment group. These data indicates that the treatment group made slightly more relevant comments on peer performance. Descriptive statistics for the overall citation frequency scores were $M = 1.19$, $SD = .48$ for the control group and $M = 1.48$, $SD = .48$ for the treatment group. Again, the treatment group seems to mention more skill aspects per comment than the control group. Both the average relevance and citation scores were subjected to a MANOVA in order to examine the group difference. The results were significant, Wilks' $\lambda = .85$ ($F(2, 50) = 4.322$), $p < .05$, $\eta^2 = .14$. A post-hoc ANOVA on each measure suggests that this group difference lies in both measures, $F(1, 51) = 5.24$, $p < .05$, $\eta^2 = .093$ for relevance scores and $F(1, 51) = 4.66$, $p < .05$, $\eta^2 = .084$ for citation scores. The results suggest that the treatment group made



Notes: Measr = Rasch logit measures; T1, T2 = Rater severity measures of Treatment 1 and 2; Inst= Rater severity measures of Instructors; Ratees = Presentation quality measures; Items = item difficulty measures.

Figure 2 Results of Rasch analysis of visual aspect

Table 2 Item difficulty and fit statistics of verbal and visual aspects

Skill	Item	ID	Fit			
			Infit	*z*	Outfit	*z*
Verbal	Vocabulary	2.04	1.4	9	1.4	9
	Language use	1.22	1.0	0	.9	−1
	Diction	−.06	.9	−2	.9	−2
	Conclusion	−.35	.9	−2	1.0	0
	Intonation	−.68	.8	−4	.9	−2
	Body	−.71	1.0	0	1.0	0
	Pace	−.73	.9	−4	.9	−1
	Introduction	−.74	.9	−3	.8	−3
Visual	Visual aids	.53	1.1	4	1.1	1
	Posture	.33	.9	−4	.9	−3
	Eye contact	−.25	1.0	0	1.0	0
	Gesture	−.61	1.0	0	1.0	0
	Facets		Separation		Reliability	
Verbal	Item		25.37		1.00	
	Rater		7.46		.98	
	Presentation		2.99		.90	
Visual	Item		11.09		.99	
	Rater		4.94		.96	
	Presentation		1.90		.78	

Notes: ID = item difficulty. The higher the item difficulty is, the more difficult it is.

more relevant comments and mentioned more skill aspects per comment than the control group.

Finally, Table 5 shows the raw frequency of skill citation and average skill frequency per individual according to category. These raw frequency results do not seem to show clear differences between the two groups, partly because the numbers of total comments used were different. To test the hypothesis, the average frequency of citation within a single peer rater was subjected to the Mann-Whitney tests. The results suggest that the treatment group made more frequent citations per comment on Language use ($U = 208.5$, $p < .01$, $r = -.43$), Conclusion ($U = 259$, $p < .01$, $r = -.34$), Gestures ($U = 112$, $p < .01$, $r = -.61$), and Visual aids ($U = 236.5$, $p < .05$, $r = -.28$). Although these differences further support the hypothesis, all actual individual's mean scores of each skill aspect turned out to be less than one, thus leaving this micro-level analysis questionable.

Granted, the first and second analyses of comments consistently provide evidence for the benefits that the treatment group seemed to

Table 3 Descriptive statistics of presentation quality and rater severity measures

Skill		Presentation quality						Rater severity			
		Treatment1		Treatment2		Control					
		Peer	Inst	Peer	Inst	Peer	Inst	T1	T2	C	Inst
Verbal	Mean	1.72	1.25	2.40	.93	2.68	.87	.65	−.25	−.54	1.22
	SD	.60	.69	.91	.81	.90	1.01	.73	.95	1.15	.47
	Min.	.78	.13	1.06	−.33	1.49	−.82	−1.79	−2.30	−2.92	.71
	Max.	3.17	2.92	4.34	2.68	4.88	3.96	1.83	1.22	.92	1.69
Visual	Mean	1.84	1.93	2.74	1.66	3.15	1.39	.93	−.22	−.82	1.10
	SD	.77	.86	.77	.82	0.91	1.03	.86	.96	1.37	.53
	Min.	.17	.18	1.50	.42	0.81	−1.42	−1.81	−2.29	−4.04	.57
	Max.	3.83	3.76	4.00	3.51	4.73	3.09	2.14	1.13	1.24	1.63

Notes: T = Treatment; C = Control; Inst = Instructor. The number of observations for each group was 26 (T1), 28 (T2), and 27 (C). There were four instructor raters. Note that the higher the measure is, the better the quality is and the more severe the rater.

Table 4 Results of regression analysis of each group on instructor rating

Skill	Model	Unstandardized coefficients		Standardized coefficients				
		B	Std. error	Beta	*t*	*p*	*r*	Adj. r^2
Verbal	Constant	−.121	.310		−.391	.699		
	Treatment 1	.796	.170	.691	4.684	.000	.691	.478
	Constant	−.652	.305		−2.135	.042		
	Treatment 2	.678	.121	.741	5.628	.000	.741	.549
	Constant	−1.528	.389		−3.923	.001		
	Control	.911	.138	.797	6.594	.000	.797	.635
Visual	Constant	.692	.363		1.905	.069		
	Treatment 1	.675	.182	.604	3.714	.001	.604	.365
	Constant	−.542	.385		−1.406	.171		
	Treatment 2	.801	.136	.756	5.896	.000	.756	.572
	Constant	−1.531	.418		−3.665	.001		
	Control	.935	.128	.826	7.313	.000	.826	.681

Notes: $n = 26$ (Treatment 1), 28 (Treatment 2), and 27 (Control).

gain from training, which bolsters the second hypothesis. That is, the treatment group produced more relevant comments and mentioned more skill aspects per comment compared to the control group. This part of the research conforms to the results of L1 and L2 writing research on peer response groups (Berg, 1999; McGroaty & Zhu, 1997; Stanley, 1992).

Finally, the present research generated an averaged adjusted r^2 of .523 (95% CI = .309 – .654, $F(1, 46) = 50.36$) from both Studies 1 and 2. These values are plotted in Figure 3, along with those of previous studies. As can be seen in Figure 3, adding the present study does not quite change the pooled average of effect size ($M = .506$, 95% CI = .416 − .586). In turn, it does not change the non-significant results of the diffuse test for the heterogeneity of effect sizes ($\chi^2 = 1.22$, $df = 4$, $p > .05$)), suggesting homogeneity of effect sizes.

One interpretation that can emanate from these data is the robustness of peer assessment, in the sense that it achieves the average effect size of .506 based on the five EFL peer assessment studies, whose adjusted r^2 values are pooled from studies with various settings, including both trained and non-trained groups in the present study. Except for the small sample in the Cheng and Warren study (1999), all other studies reached averaged effect sizes of above .50. Such an effect size is almost twice as large as .26, which was what

Table 5 Frequency counts of comment and mean citation frequency by category

Skill	Category	Counts (Mean)	
		Control	Treatment
Verbal	Intonation	8 (.03)	4 (.00)
	Diction	79 (.19)	46 (.15)
	Pace	78 (.15)	67 (.13)
	Language use	5 (.00)	23 (.05)
	Vocabulary	29 (.04)	51 (.02)
	Introduction	3 (.00)	4 (.00)
	Body	88 (.20)	120 (.22)
	Conclusion	5 (.00)	9 (.01)
	Pronunciation*	60 (.13)	73 (.14)
	Length*	8 (.00)	5 (.01)
Visual	Eye contact	42 (.07)	38 (.10)
	Gesture	22 (.04)	95 (.17)
	Posture	20 (.05)	10 (.02)
	Visual aid	135 (.29)	192 (.38)

Notes: * = categories added by raters.
n = 27 (control), 26 (treatment). The number in the parentheses are mean citation frequency per single peer rater.

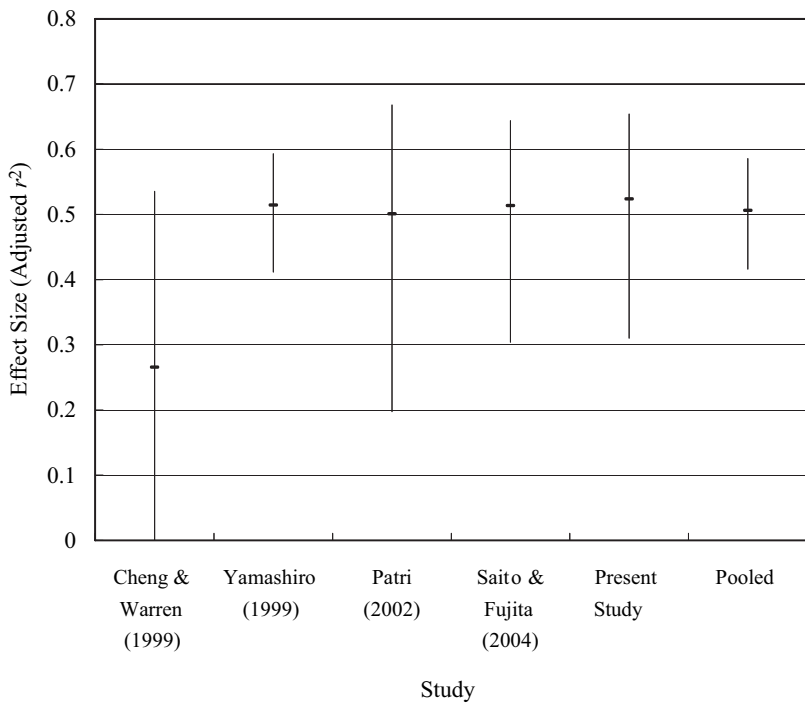


Figure 3 Effect sizes and 95% confidence intervals of five EFL peer assessment studies

Cohen (1988, p. 414) conceived to be a large effect size for r^2 in behavioral sciences in general. In fact, a meta-analysis (Falchikov & Goldfinch, 2000) of teacher–peer assessment correlation retrieved from 56 studies in the general education literature generated a mean r of .69 (roughly equal to .47 when squared), which is similar to the present study. This preliminary summary supports the robustness of peer assessment in EFL classrooms, although the number of studies included here is still not large enough and there is a possibility of publication bias.

IV General discussion and conclusion

In the present study, two hypotheses were investigated. The first hypothesis was not supported in the short or long training conditions. There are several possible reasons for this result. First, both groups receiving instruction on all of the oral presentation skill aspects may have established a certain level of correlation between the instructors and students, and training may not have added any more advantages to the treatment group on top of instruction in the skill aspects. However, before this interpretation is accepted, comparisons with a control group receiving no instruction are necessary.

A second speculation related to the null results of the second study comes from the intrinsic characteristics of each intact group. Although placement test scores and the *post hoc* comparison of presentation quality indicated that there were no particular proficiency advantages for any group, there is still a possibility that either the differences in major (see Falchikov & Goldfinch, 2000) or class specific characteristics affected the strength of the correlations between peer and instructor ratings, as in Cheng and Warren (1999). This intrinsic difference might have washed away any advantages of training.

The comment data of the present study support the second hypothesis because of the significant difference in relevance judgments and frequency of citation on overall skill aspects. Several interesting pictures emerge from the results. First, the present results suggest that instruction on skill aspects without rater training may be sufficient for peer assessment to correlate with instructor scores to a certain degree; however, rater training seems to help students provide more relevant comments, citing more skill aspects on peer performance. Second, rater training reduces the possibility of a misfit in the data. In Study 1, all three instructors were misfitting because the frame of reference in the data set by peer data, representing the majority, differs from that

of instructors. Longer training in Study 2 seems to have converged the frame of reference of learners with that of instructors, which has reduced the number of misfitting peer raters and eliminates misfitting instructors. From these two observations, the conclusion is made that rater training does not statistically improve the correlation with instructors, but it does have an effect at a different level of rating behavior, such as commenting and setting a similar frame of reference. However, the present data cannot answer the question as to who has actually benefited from the longer training. Could it be peer learners alone or instructors? Or both groups?

The results of the present study should be interpreted with some caution, due to some methodological limitations. The first limitation of the present study is that it did not include a pretest. Although the questionnaire used in the study confirmed the students' lack of experience in peer assessment, the absence of a pretest may make uncertain any changes that may result from rater training. There is a problem, however, if the control group experiences any peer rating as a pretest that can be considered training. In this study, the control groups truly had no prior experience in peer assessment.

Another limitation of this study is the shortage of training time. It can be argued that neither the short or long training sessions employed here were sufficient for establishing strong correlations among raters. In their discussion of the use of a pre-listening activity, however, some researchers (e.g. Field, 2002) have cautioned that spending too much time on a pre-listening activity reduces actual listening time and, hence, learning time. The same argument may be applicable for training time for rating in classrooms. Rating itself is not an essential skill in language learning, and rater training should not take up too much time in language instruction, although having an explicit knowledge of how to discern a good performance may improve a student's own performance. Training time used for this study reflects, to a certain extent, the amount of time a real classroom should afford.

When considering the application of the present results, a number of issue remain unresolved. Whether long training and peer assessment have pedagogical value in language classrooms is an issue to explore in further studies. In particular, the connection between these tasks and an actual improvement in language performance remains unclear. Long training for better comments may not be worth the time unless such commenting skills assure an improvement in student language performance. A connection between peer response training and improvement in the student product is examined in writing research, but the results have not been consistent (Berg, 1999;

McGroarty & Zhu, 1997). Although the present study has shown the effects of training on comments, in that training leads to better comments, the actual effect size ($\eta^2 = .14$) may still not be large enough for practical significance.

Despite these limitations, the results of the present study support the following conclusions. First, when instruction on skill aspects is given, this may not, contrary to the expectation, result in better correlations between learner and instructor ratings. Peer assessment is fairly robust (reliable without much training), and longer training alone may not provide further improvement in correlation. On the other hand, longer training may reduce the possibility of instructor misfit because of improvement in the overall frame of reference in either the majority peer rating, instructor rating, or both peer and instructor ratings. Second, rater training for peer assessment may raise language learners' awareness of skill aspects and lead to enhancement in the frequency and relevance of comments. In this sense, peer assessment training is a meta-cognitive activity in which student attention is drawn to the features of a language learning task. If increasing consciousness of the performance criteria is the prerequisite for learning, then peer assessment training may facilitate the process (Black & Wiliam, 1998).

Acknowledgements

I am grateful for all the comments by the two reviewers of *Language Testing* and encouragements from the Editor of *Language Testing*. This research was made possible through a grant by the Japan Ministry of Education, Culture, Sports, Science, & Technology (task no. 19520475).

V References

Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing, 8*, 225–241.

Black, P., & **Wiliam, D.** (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–68.

Cheng, W., & **Warren, M.** (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*, 233–239.

Cheng, W., & **Warren, M.** (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education, 24*, 301–314.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51.

Devenney, R. (1989). How ESL teachers and peers evaluate and respond to student writing. *RELC Journal, 20*, 77–90.

DiPardo, A., & **Freedman, S. W.** (1988). Peer response groups in the writing classroom: Theoretic foundations and new directions. *Review of Educational Research, 58*, 119–149.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment & Evaluation in Higher Education, 11*, 146–165.

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

Field, J. (2002). The changing face of listening. In J. C. Richards & W. A. Renandya (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 242–247). Cambridge, UK: Cambridge University Press.

Fletcher, C., & **Baldry, C.** (1999). Multi-source feedback systems: A research perspective. *International Review of Industrial and Organizational Psychology, 14*, 149–193.

Haaga, D. A. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*, 28–32.

Harris, M. M., & **Schaubroeck, J.** (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43–62.

Hu, G. (2005). Using peer review with Chinese ESL student writers. *Language Teaching Research, 9*, 321–342.

Jafarpur, A. (1991). Can naive EFL learners estimate their own proficiency? *Evaluation and Research in Education, 5*, 145–157.

Liu, J., & **Hansen, J. G.** (2002). *Peer response in second language writing classrooms*. Ann Arbor, MI: The University of Michigan Press.

Long, M. H., & **Porter, P. A.** (1985). Group work, interlanguage talk, and second language learning. *TESOL Quarterly, 19*, 207–228.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246–276.

Lumley, T., & **McNamara, T. F.** (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54–71.

McGroarty, M. E., & **Zhu, W.** (1997). Triangulation in classroom research: A study of peer revision. *Language Learning, 47*, 1–43.

McNamara, T., & **Lumley, T.** (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–156.

Morahan-Martin, J. (1996). Should peers' evaluations be used in class projects? Questions regarding reliability, leniency, and acceptance. *Psychological Reports, 78*, 1243–1250.

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Thousand Oaks, CA: Sage.

Rothschild, D., & **Klingenberg, F.** (1990). Self and peer evaluation of writing in the interactive ESL classroom: An exploratory study. *TESL Canada Journal, 8*, 52–65.

Saito, H. (1999). Dependence and interaction in frequency data analysis in SLA research. *Studies in Second Language Acquisition, 21*, 453–476.

Saito, H., & **Fujita, T.** (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*, 31–54.

Schoonen, R., Vergeer, M. M., & **Eiting, M.** (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*, 157–184.

Shohamy, E., Gordon, C. M., & **Kraemer, R.** (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*, 27–33.

Smithson, M. J. (n.d).: *Scripts and software for noncentral confidence interval and power calculations: noncentral F files for SPSS*. Retrieved December 20, 2005, from http://psychology.anu.edu.au/ people/smithson/ details/CIstuff/CI.html

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment & Evaluation in Higher Education, 18*, 221–233.

Stanley, J. (1992). Coaching student writers to be effective peer evaluators. *Journal of Second Language Writing, 1*, 217–233.

Stefani, L. A. (1992). Comparison of collaborative self, peer and tutor assessment in a biochemistry practical. *Biochemical Education, 20*, 148–151.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25–32.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276.

Tornow, W. W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management, 32*, 221–229.

Webb, N. M. (1982). Student interaction and learning in small groups. *Review of Educational Research, 52*(3), 421–445.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.

Wright, B., & **Linacre, J. M.** (1994). Reasonable mean-square fit values. *Rasch Measurement: Transaction of the Rasch Measurement SIG, 8*, 370.

Yamashiro, A. (1999). Using structural equation modeling to validate a rating scale. Presented at the 21st Language Testing Research Colloquium, Tsukuba, Japan.

Yamashiro, A. D., & **Johnson, J.** (1997). Public speaking in EFL: Elements for course design. *The Language Teacher, 21*, 13–17.

Appendix 1: Schedule for instruction and training

Class	Classroom activities (focused skill aspects for oral presentation)	Brief description	Treatment in Study 1	Treatment in Study 2
1	Information gap (Gestures/Posture) Questionnaire on peer-assessment experience	1) Students are paired and guess the meanings of each other's gestures. 2) They stand up and introduce themselves in pairs, paying attention to postures.		✓
2	Group mock presentation (Eye contact/ Visual aids)	Students are assigned to groups, with each group member being assigned a short text of different themes (based on a worksheet). Each student makes a visual aid to explain his or her theme and presents it within the group, while paying attention to eye contact.		✓
3	Group mock presentation (Introduction/ Body/Conclusion)	Each student makes a brief speech draft for "my hometown" using a template focusing on the organization of the introduction, body and conclusion. Each presents it in the group.		✓
4	Class mock presentation (Diction/Pace/ Intonation)	In the 4th and 5th sessions, all students are asked to bring in a short memo that includes their main topic and the subtopics.		✓
5	Class mock presentation (Grammar/ Vocabulary)	They introduce themselves as well as their topics and subtopics for a public speech in front of the class, while paying attention to the focused skills.	✓	✓
6–8	Presentations and peer assessment			
9	Feedback	Students receive comments from peers and instructor.		

Note: ✓ = Treatment groups experienced rater training.

Appendix 2: Rubric for classroom peer assessment of oral presentation. (Adapted from Yamashiro and Johnson, 1997)

	Skill aspect items	Superior (4)	Adequate (3)	Minimal (2)	Need work (1)
Visual skills (physical)	Posture (Standing with back straight and looking relaxed)	Back straightened. Looks comfortable.	Moderate posture.	Some problems with posture.	Sways or fidgets all the time. Looks uncomfortable.
	Eye contact (Looking each audience member in the eye)	Continuous eye contact. Rarely refers to notes.	Moderate eye contact. Occasional reference to notes	Limited eye contact. Frequent reference to notes.	No eye contact
	Gesture (Using some, well-timed gestures, nothing distracting)	Uses gestures and expressions to enhance the presentation. Natural.	Occasional use of hands and body movement. Sometimes effective.	Ineffective. Rarely used	Distracting. Or no gestures
(visual)	Visual aids (Using visual aids effectively)	Effective use of visual aids.	Effective to some extent.	Not so effective. Rarely used	Ineffective or no use.
Verbal skills (organization and content)	Introduction (Introducing thesis statement, sub-topics)	Main theme is clearly delineated, and all the sub-topics are listed.	Main theme and sub-topics delineated well to a certain degree.	Main theme and sub-topics delineated insufficiently or briefly.	No introduction.
	Body (Presentation of details of main themes and subtopics with attractive content)	Details are explained. All the sub-topics are covered. The content is attractive.	Details are explained. All the sub-topics are covered to a certain degree.	Brief, insufficient presentation of details.	Problems with content. No clear main point. Not organized well.
	Conclusion (Including restatement/ summation and closing statement)	Restatement of major topics and concluding remarks provided.	Major topics summarized. Concluding remarks missed.	Brief, insufficient summary of major topics.	No conclusion

(delivery)	Pace (Speaking at a good rate – not too fast, not too slow – with appropriate pauses)	Fluid, natural delivery. Appropriate pauses.	Adequate pace. A few longer pauses.	Long pauses at several places. Some unevenness of pace.	Halting, uneven pace. Distracting.
	Intonation (Speaking using proper pitch patterns)	Adequate intonation throughout.	Mostly adequate, but some indication of unnaturalness	Many inadequate intonations.	Unnatural, strange intonation throughout.
	Diction (Speaking clearly – no mumbling or interfering accent)	Clear articulation all the time.	Adequate articulation. Mostly clear.	Some unclearness.	Mumbling. Unclear.
(language)	Language Use (Using clear and correct sentence forms)	Grammatical and fully comprehensible.	A few local errors but do not affect comprehension.	Some global errors affect comprehensibility.	Numerous errors. Difficult to comprehend.
	Vocabulary (Using vocabulary appropriate to the audience)	Use of adequate vocabulary. Variety.	Used a few inadequate vocabulary terms.	Some vocabulary inadequacy. Limited vocabulary.	Numerous instances of inadequate vocabulary use. Very limited vocabulary