The influence of peer feedback on self- and peer-assessment of oral skills

Mrudula Patri *City University of Hong Kong*

This article investigates the agreement amongst teacher-, self- and peer-assessments of students in the presence of peer feedback. This is done in the context of oral presentation skills of first year undergraduate students of ethnic Chinese background. The research instrument consisted of a self- and peer-assessment questionnaire containing 14 items related to the organization of the presentation content, use of language and interaction with the audience. The participants had taken part in a training and practice session on self- and peer-assessment before engaging in the assessment tasks. The findings show that, when assessment criteria are firmly set, peer-feedback enables students to judge the performance of their peers in a manner comparable to those of the teachers. However, the same is not found to be true with self-assessment.

I Introduction

This article examines the effectiveness of self- and peer-assessment augmented by peer-feedback while testing oral presentation skills. The focus is on training within the limits imposed by practical classroom considerations and the effectiveness of self- and peer-assessments. If the effectiveness of self- and peer-assessment could be adequately improved, the teachers' workload could be partly reduced. Teachers could then focus more on enhancing their teaching techniques.

1 Validity of self- and peer-assessments

Self- and peer-assessments have gained much attention in recent years owing to growing emphasis on learner independence and learner autonomy. Further, self-assessment, peer-involvement and peer-assessment have been viewed as having significant pedagogic value. However, studies on the validity of self- and peer-assessments have revealed some contradictory results. Bachman and Palmer (1989)

Address for correspondence: Mrudula Patri, Lecturer, English Language Centre, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; email: elpatri@cityu.edu.hk

Language Testing 2002 19 (2) 109–131 10.1191/0265532202lt224oa

believe self-assessment to be a valid and reliable measure of communicative language ability. Williams (1992) reports close agreement between self-ratings and teacher ratings when the latter are used as a reference. Stefani (1994) analysed the correlation between self- and tutor-assessment and found that tutor's marks closely matched students' self-assessments (r-value = 0.93). Other studies showing high teacher–self correlations include Oldfield and Macalpine (1985) and Sullivan and Hall (1997). On the other hand, some studies recorded low agreement between the two (Jafarpur, 1991; Hughes and Large, 1993; Mowl and Pain, 1995; Orsmond *et al*., 1997).

With regard to peer-assessment, studies by Rolfe (1990), Hughes and Large (1993), Miller and Ng (1994) and Freeman (1995) have noted high agreement between teacher- and peer-assessments. Jafarpur (1991), Falchikov (1995), Mowl and Pain (1995), Kwan and Leung (1996), and Orsmond *et al*. (1997) observed otherwise.

Previous studies on self- and peer-assessments found that learners over- or under-estimating their own and their peers' language skills affects the validity of assessments (Boud and Tyree, 1979; Wangsotorn, 1980; Armanet and Obese-jecty, 1981; Heilenmann, 1990; Rolfe, 1990) with low achievers over-estimating and high achievers under-estimating (Stefani, 1994; Falchikov, 1995; Mowl and Pain, 1995; Kwan and Leung, 1996; Orsmond, *et al*., 1997). Pond *et al*. (1995) defined over-marking by peers as 'friendship marking' or 'decibel marking'. This could be because peers find it difficult to criticize their friends (Falchikov, 1995).

Others observed that learners' ability to respond depends on their willingness to answer the questions posed (Heilenmann, 1990), on the choice of descriptors used in the rating scale (Davidson and Henning, 1985) and on their ability to understand the questionnaire items (Heilenmann, 1990). Shore *et al*. (1992) noted that learners find inferential dimensions, such as appropriacy and fluency, more difficult to assess than performance dimensions. In addition, less able students find it more difficult to self- or peer-assess when compared to more able students (Jafarpur, 1991; Pond *et al*., 1995; Orsmond *et al*., 1997; Sullivan and Hall, 1997).

Marking is a subjective activity. Orsmond *et al*. (1997) noted that subjectivity may apply both to self- and peer-assessment practices. Therefore more guidance on the marking criteria should be given to ensure that all markers can apply previously agreed criteria in a consistent fashion (Sullivan and Hall, 1997; Woolhouse, 1999). Orsmond *et al*. (1997) point out that clear marking criteria give students the opportunity to see how their marks have been calculated. However, they caution that there could be differences in students' estimations

as to how well they have performed regardless of the fact that they understand the marking criteria.

Referring to oral testing, Underhill (1987: 3–5) commented that a greater degree of subjectivity could be involved in oral testing because, in an oral test, what is assessed is the oral 'message' that is communicated between people or conveyed by the speaker. Emphasizing the same, Freeman (1995) adds that students must be given adequate training and practice in peer-assessment in order to minimize potential inconsistencies associated with subjectivity. He recommends showing videos and drawing students' attention to elements of best and worst presentations. Learner training and guidance in interpreting the marking criteria was also emphasized by Oldfield and Macalpine (1995), Jafarpur (1991), Adams and King (1995), Mowl and Pain (1995), Pond *et al*. (1995), Kwan and Leung (1996), and Orsmond *et al*. (1997).

To increase the validity of self-assessment, some suggested the use of task-linked questionnaires (Oskarsson, 1984; Cameron, 1990). With regard to peer-assessment, Falchikov (1995) found 'Peer Feedback Marking' on students of Human Developmental Psychology to be useful. Freeman (1995) found that discussions amongst peers before awarding a team rating for oral presentations leads to closer agreements between staff and student ratings.

It is apparent that peer-involvement creates opportunities for interaction while increasing objectivity in assessment. If learners are placed in a situation where they can access information on the quality and level of their own performances or those of their peers, then it is possible that they will be able to clarify their own understanding of the assessment criteria (either set by students themselves or by the teacher) and, more importantly, what is required of them.

Following the above observations related to peer-feedback and peer-involvement, is it possible that:

> If peer discussion and feedback on a student's presentation before giving a mark can lead to a closer agreement between teacher and peer-assessment, will it also have a similar effect on teacher and self-assessments?

This article attempts to find the answer by investigating the influence of oral feedback on self- and peer-assessment of individual oral presentations. This is done after completing a two-hour training session followed by practice at the rate of two hours per week over a period of four weeks. The participants were first year undergraduate students of ethnic Chinese background.

2 Aims

The experiments conducted in the present study were designed to test the following two hypotheses in the context of assessing oral presentation skills. When the assessment criteria are firmly set:

1) peer feedback will enable students to produce judgements about themselves that are comparable to those made by the teacher;
2) peer feedback will enable students to produce judgements about their peers that are comparable to those made by the teacher.

These hypotheses were tested mainly by correlating self- and peer-assessments with those of the teacher in the context of assessing students' oral presentations. The same questionnaire was used for teacher-, peer- and self-assessments with the minimum possible modifications to ensure uniformity of assessment. In order to isolate the influence of peer feedback, the same procedure was applied both to control groups and experimental groups drawn from an apparently homogenous student population. Peer feedback was absent in the case of the control group whereas it was introduced in a structured manner into the experimental group.

II Method

1 Participants

The participants in the study included 56 native Chinese students aged between 18 and 21 years, from the City University of Hong Kong (Table 1). Forty-one of the participants were first year students of Bachelor of Science in Computer Mathematics and Information Systems attending the 'English Foundation Programme' which ran over

Table 1 Organization of participants in the control and experimental groups

Group	Class	Course taught by the researcher	Program at City University	Number of students
Control	A	EFP	BSc Computer Mathematics	11
Control	B	Special	BA Business Studies, Accountancy	8
Control	C	EFP	BSc Information Systems	11
Experimental	A	EFP	BSc Computer Mathematics	9
Experimental	B	EFP	BSc Computer Mathematics	10
Experimental	C	Special	BA Business Studies, Accountancy	7

a period of 14 weeks. This was a mandatory remedial English program for students who had obtained a 'D' or 'E' grade in the 'Use of English' examination of the Hong Kong Advanced Level Examinations (HKALE).[1] The rest were degree students from Business Studies and Accountancy and were attending a special course entitled Practice Speaking for Communication. (Both the courses were taught by the author.)

To evaluate the effect of peer feedback on self- and peer-assessments, control and experimental groups were set up (Table 1). Participants in the former group performed the tasks of self- and peer-assessment complemented with peer feedback while the latter did so without any peer feedback.

The control group consisted of 11 students of Computer Mathematics (class A), 8 students of Business Studies and Accountancy (class B) and 11 students of Information Systems (class C), respectively. The experimental group included classes A, B and C with 19 students of Computer Mathematics (classes A and B) and 7 students of Business Studies and Accountancy (class C).

2 *Oral presentation task*

a English Foundation Programme The English Foundation Programme (EFP) was a 28-hour remedial English course with two hours of instruction per week. The course focused on improving the students' reading, speaking, listening and writing skills. Coursework assessment included one Reading, one Speaking, one Listening and one Writing test. With regard to speaking skills, the students were required to make oral presentations on a topic given by the teacher. Their performances were assessed for organization of content, language use, manner and interaction with the audience. The topics chosen were based on students' perceived ability, familiarity with the topic and interest. While preparing for the assessment task, the students were given practice in making an oral presentation and the teacher was expected to give feedback on their performance. The experiment was built into the teacher's lesson plan.

b Practice Speaking for Communication 'Practice Speaking for Communication' was an 'out-of-discipline' course offered by the Language Institute of City University of Hong Kong. It was open to all students of the University. Enrolment was limited to 20 students on

[1]HKALE is the examination that Form 7 students sit for before entering the University. Grades D and E are approximately equal to TOEFL scores of 503 and 541 respectively (Hong Kong Examinations Authority, 1990).

a first-come-first-served basis. Similar to the EFP, the course was 28 hours long with two hours of instruction each week. The course curriculum (with no assessment component) included making oral presentations with a focus on organization of the talk, language use, manner and interaction with the audience. The experiment conducted was built into the teacher's lesson plan.

3 Research instrument: Self-assessment questionnaire (Appendix 1)

The design of the self-assessment questionnaire was based on the course objectives (as explained previously) where students' presentations were assessed for organization of the presentation content, language use, manner and interaction with the audience. Fourteen questions were included in the questionnaire. The questions were simple and direct in wording (as suggested by Underhill, 1987) so as to elicit direct responses from students. The questions were divided into four categories. Category 1 included questions 1–6 that were related to organization and content of presentation. Category 2 (questions 7–9) was based on the use of language. Category 3 (questions 10–12) was related to manner. Finally, Category 4 included questions 13 and 14 which were related to interaction with the audience. The questions were phrased so as to develop students' understanding of the assessment task (Orsmond *et al.*, 1997) and work towards those objectives. The rating scale was based on a 5-point Likert scale so that each assessor would categorize performance as being: 1 – poor; 2 – unsatisfactory; 3 – satisfactory; 4 – good; 5 – excellent.

The same questionnaire was used for teacher-, peer- and self-assessments with minimal modifications; changes were limited to wording in the rubrics:

- self-assessment questionnaire: 'Rate yourself using the scale';
- peer-assessment questionnaire: 'Rate your classmate using the scale'; and
- teacher assessment questionnaire: 'Rate the student using the scale'.

4 Training session

The training session lasted for about two hours of class time. The main purpose of the training session was to establish the assessment criteria. During the training session, the students (both from the control and experimental groups) were first given a worksheet to introduce them to the important elements of a good presentation. The worksheet focused on the presentation format, content and language (presented as useful expressions) and suggestions on delivering presentations (facial expressions, movements and gestures, eye contact,

disadvantages of speaking from notes, etc.). Participants were given 15 minutes to familiarize themselves with the contents of the worksheet. The class was then divided into two groups. One group was given the topic 'Chinese New Year holidays are the most enjoyable' and the other group was given the topic 'Chinese New Year holidays are the most stressful'. Participants were then asked to spend 30 minutes preparing for their presentations. With a view to familiarizing the participants with the marking criteria, at the end of 30 minutes the researcher elaborated on the questionnaire items on the peer- and self-assessment forms by explaining what the participants should focus on while assessing their own and their peers' talks. Questionnaire items 1–7 (eliciting responses on the organization and content) were explained in relation to the performances after watching a sample video (explained below) consisting of a good, an average and a poor presentation (the evaluations 'good', 'average' and 'poor' were made in relation to the criteria set by the teacher as explained below). However, items 8–14 – which referred to fluency, pronunciation, confidence, speaker's eye contact, verbal and non-verbal interaction – had to be explained in more detail as:

- Fluency (point 8 in the questionnaire): Is the student pausing in the middle of his/her presentations because s/he cannot think of the right word or does not know what to say?
- Pronunciation (point 9): Does the student have problems pronouncing even the most common words?[2]
- Confidence (point 10): Is the student nervous?
- Confidence (point 11): Is the student looking at his/her notes all the time or simply reading aloud?
- Eye contact (point 12): Is the student looking at the ceiling or the floor? Is s/he looking at only one person?
- Non-verbal communication (point 13): Is the student using hand gestures as a natural means of conveying his/her message? Does the student maintain a pleasant facial expression?
- Verbal communication (point 14): Is the student involving the audience in the presentation by asking questions? Are the members actively participating in the presentation?

A video-tape including three student presentations made by EFP students from the previous semester was used (similar to the procedure recommended in Freeman, 1995) to clearly establish the criteria set by the researcher. The researcher had assessed the students in the video prior to the training session. Students assessed these by filling

[2]'Common' in 'common words' is in relation to the vocabulary items tertiary students in a remedial ESL class are expected to be familiar with.

in the peer-assessment forms. To firmly set the assessment criteria, the video was stopped and the teacher and the participants compared their assessments at the end of each performance. When there were differences, the researcher explained the basis for her judgements. For example, 'this student has provided a lot of detail and the detail is relevant and varied. That is why I gave him a "4".' Every training session (three sessions, i.e. classes A, B and C for the control group, and classes A, B and C for the experimental group) was audio-taped. Prior to each session, the researcher checked the procedures followed during the previous session so as to standardize the training sessions.

III Procedure

In week 1, the week after the training session, participants from both the control and experimental groups made their presentations. Half the class presented topic (a): 'Chinese should be the medium of instruction in secondary schools' and the other half topic (b): 'Comics have a bad influence on school children.'

1 Week 1: Experimental group

After the training session, participants made their oral presentations on the two topics. Participants were divided into groups of three or four depending on the class size. Each group was asked to assess their peers. Group members were given a self-assessment form and an appropriate number (equal to the number of students present in the group) of peer-assessment and peer-feedback (PF) forms (see Appendix 2). Participants then entered the names of their group members on both the peer-assessment and PF forms. Each student from the class made a three-minute presentation on the topic assigned by the researcher. The researcher assessed the participants during the presentation using the teacher-assessment form. During the presentation, peers noted their comments on the PF forms. Following the presentation, participants sat in groups of three or four and commented on their peers' talks. The researcher spent about five minutes with each group noting down the comments made by the peers during the feedback session. After each feedback session, each member of the group filled in the self-assessment form and then filled in the peer-assessment form. (Participants referred to the PF form during peer-assessment; see Figure 1.) No feedback was given by the researcher on the individuals' talks. At the end of each session, the researcher collected the self- and peer-assessment questionnaires and the PF forms. The participants were given two new topics to prepare for their presentations in the following week.

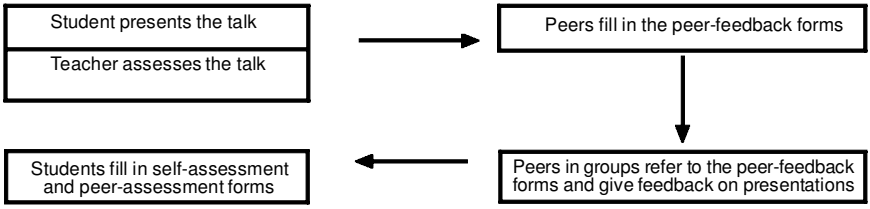

Figure 1 Self- and peer-assessment in the experimental group

2 Week 1: Control group

The participants in the control group followed the same procedure as the experimental group, but there was no peer feedback after the talks (Figure 2). At the end of the session, participants were given two new topics to present in the following week. Both groups followed the above procedure during weeks 2–5.

Participants were placed in different groups every week to encourage maximum interaction and to expose students to different points of view. Table 2 shows the grouping of students. For example, in week 1, group one is made up of students A, B and C, group 2 includes students D, E and F and group 3 consists of students G, H and I. In week 2 they are re-arranged to have students A, E and H in group 1, students D, B and I in group 2 and G, C and F in group 3. After the session in week 5, the researcher collected the self- and

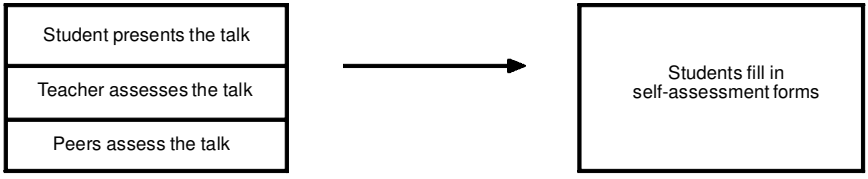

Figure 2 Self- and peer-assessment in the control group

Table 2 Peer feedback sessions

Peer-feedback session	Group 1 students	Group 2 students	Group 3 students
Week 1	ABC	DEF	GHI
Week 2	AEH	DBI	GCF
Week 3	AFI	DCH	BEG
Week 4	ABC	DEF	GHI
Week 5	AEH	DBI	GCF

Note: Letters A–I do not correspond to any particular student or class. They are used to illustrate the change in grouping of students from weeks 1–5.

peer-assessment forms as well as the peer-feedback forms. Next, the participants filled the Self- and Peer- Assessment Evaluation Forms (Appendices 3 and 4).

Data obtained from week 5 was used to correlate teacher–self and teacher–peer assessments to be reported and discussed in this study. Weeks 1–4 provided participants with practice in self- and peer-assessments.

3 Topics for weeks 2–5

Since the participants were from a remedial English class and since the experiment was built into the lesson plan (refer to the section on Participants), the topics chosen were based on familiarity and general interest.

- Week 2
  a) Males are generally stronger and more competitive than females.
  b) Women with small children can/cannot work outside home.
- Week 3
  a) Reading comics can have a bad influence on children's behaviour.
  b) Raising children is as much a father's responsibility as is mother's.
- Week 4
  a) It is/isn't a good idea for students to have part-time jobs.
  b) It is/isn't a good idea for Hong Kong government to introduce breath tests for possibly drunken drivers.
- Week 5
  a) The Hong Kong government should/should not import labour from China after 1997.
  b) Capital punishment should not be implemented in Hong Kong.

IV Results

The data recorded included teacher-, self- and peer-assessments (TA, SA and PA) obtained in week 5 from the control and experimental groups. The number of participants dropped by natural attrition from 56 to 54 by week 5 of the investigation. (Note that students enrolled in the 'Practice Speaking for Communication' class were present on a voluntary basis and the course had no assessment component. This might have been the reason for the drop in class size.) The analysis of the data included two viewpoints: the aggregate level (i.e., at the level of a group of students) and the individual student level.

Table 3 Overall group level average ratings and standard deviations calculated from teacher-assessment, self-assessment and peer-assessment of the control and experimental groups

Asessment mode	Control group ($n = 29$)		Experimental group ($n = 25$)	
	Average	Standard deviation	Average	Standard deviation
Teacher-rating	2.71	0.33	2.89	0.34
Student's self-rating	2.87	0.43	2.98	0.4
Average of ratings by peers[a] in the class	2.94	0.26	3.08	0.28

Notes: [a]Average of ratings by peers in the class; number of peers in a class = 5 to 10.

1 Aggregate level analysis

For each student, the averages of ratings corresponding to questions 1 to 14 obtained through TA, SA and PA were estimated. The individual student average ratings were then aggregated over the corresponding group (control or experimental). The resulting group averages and the corresponding standard deviations are listed in Table 3.

Table 3 exhibits some differences in the group averages obtained through TA, SA and PA ratings for the control and experimental groups. Student *t*-tests, with alpha levels of 0.05 and 0.01 were used to test whether there was a significant difference between the average values of each pair of samples drawn from populations with the same variance. The results are presented in Table 4.

Looking at the values in Table 4, the differences between the group means for the control and experimental groups were not significant

Table 4 Aggregate level analysis: *t*-test results

Comparison	Assessment type	*t*-ratio	df	The difference beween group means at:	
				$p \leq .05$ is:	$p \leq .01$ is:
CG vs. EG	TA	2.02	52	significant	not significant
	SA	1.07	52	not significant	not significant
	PA	1.92	52	significant	not significant
SA vs. TA	CG	1.60	56	not significant	not significant
	EG	0.85	48	not significant	not significant
PA vs. TA	CG	2.98	56	significant	significant
	EG	2.14	48	significant	not significant

at the $p \leq .01$ level for TA, SA as well as PA, although differences with respect to TA and PA were significant at the $p \leq .05$ level. This is not surprising considering that both groups had consisted of students drawn from populations with similar backgrounds and language proficiency. It can also be seen that there was no significant difference between the mean scores given by TA and SA in either the absence (control group) or the presence (experimental group) of peer feedback. The situation is somewhat different when we examine the degree of similarity between PA and TA, where the difference between the averages of PA and TA was highly significant ($p \leq .01$) for the control group and significant ($p \leq .05$) for the experimental group. This suggests that peer feedback did have some effect on the ratings given by peers.

Clearly, the only substantive conclusions that can be drawn from the above aggregate level analysis are the following:

- The behaviour of PA has been somewhat different from that of SA.
- There appears to be some difference between the behaviours of PA in the presence and absence of peer feedback.

In order to obtain a better understanding of the nature of these differences, an individual level analysis was performed.

2 Analysis at the individual student level

Owing to its aggregate nature, the analysis presented in Section 1 above was unable to throw any light on the ability of a specific assessment medium to achieve discrimination. There is much research evidence (this point will be explained below in greater detail) that students performing self-assessment often tend to overrate low performance and underrate high performance. Thus, further statistical measures were computed to determine the cancelling effect and to explicitly retain the information concerning discrimination amongst individual student performances.

Pearson correlation coefficients were calculated between teacher–self (T–S) and teacher–peer (T–P) assessments for (1) each student's average rating with the teacher's average rating for that student; and (2) the mean peer rating of each student with the teacher's average rating for the student.

From Table 5 it can be seen that the T–S correlations for the control and experimental groups ($r = 0.50$, $p \leq .005$ and 0.46, $p \leq .01$, respectively) are considerably smaller than the typical r-values noted in Stefani (1994; $r = 0.93$) with respect to testing English communication skills. The low T–S correlations in the present study suggest

Table 5 Teacher–self (TS) and teacher–peer (TP) correlations: control and experimental groups

Assessment	Correlation coefficient (r)	
	Control group	Experimental group
T–S	0.50**	0.46*
T–P	0.49**	0.85**

**$p \leq 0.005$; *$p \leq .01$.

that the participants were not able to judge their performances in the same way as the teacher had done either in the presence or absence of peer feedback.

The significantly high T–P correlation ($p \leq .005$) noted for the experimental group ($r = 0.85$) in Table 5 shows that peer-assessment in this group was in high agreement with the teacher-assessment. This suggests that, in the presence of peer feedback, the students were able to make judgements of their peers' oral presentations comparable to those made by the teacher.

To examine the regression relationships implicit in the data, initially, both nonlinear (quadratic) and linear regressions were tried. However, a comparison of the two trend lines showed that they were practically identical (the deviation was no more than 3%) in every case. Therefore, linear regression was used to interpret the data. The results of the three linear regressions are presented in Figures 3a to 3d.

If there were a perfect positive linear relationship, or a correlation of +1.00 between the different pairs of ratings, the slope of the regression line would be equal to 1 and the intercept equal to 0. Thus, what the regression line can show that is missed in a single correlation coefficient is the extent to which the differences in the relationship are in the slopes of the lines or in their intercepts. Looking at the lines with T–S correlations (Figures 3a and 3b) it can be seen that both the control and experimental groups have nearly similar slopes (slope = 0.64 and 0.55, respectively) and intercepts (1.14 and 1.4) indicating only a marginal difference in their ability to assess themselves in the presence or absence of peer feedback. This further supports the observation that the participants have not been able to make judgements about their performances comparable to those made by the teacher. In particular, the fact that the regression lines and slopes were substantially smaller than 1 suggests that low achievers were over-estimating and high achievers were underestimating their performances. This finding matches that observed by Orsmond *et al.*

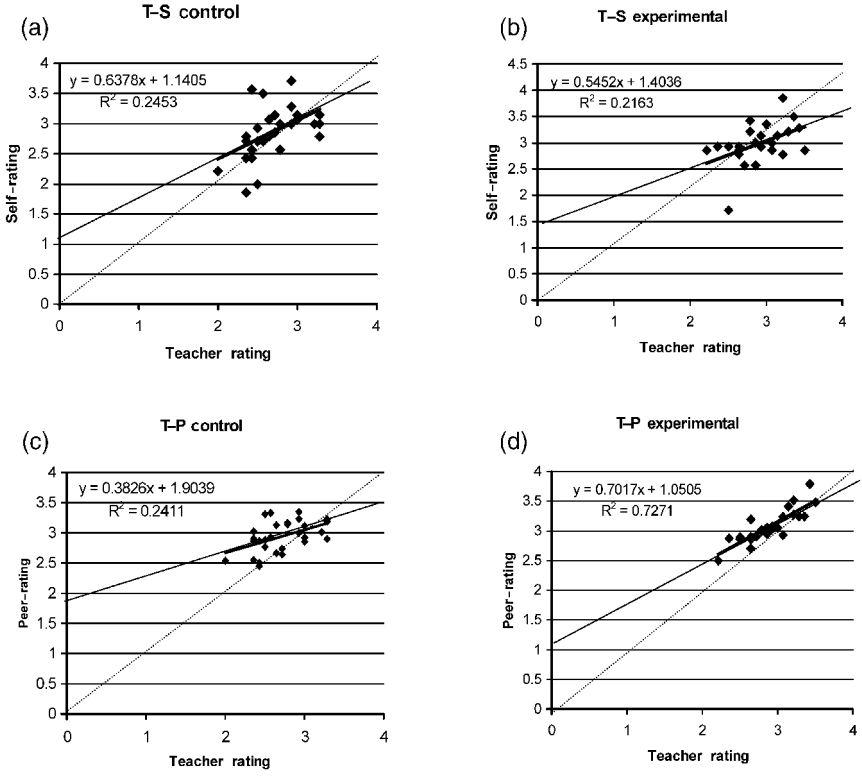

Figure 3 Teacher-, self- and peer-ratings in the control and experimental groups

(1997) and Stefani (1994). However, it is possible that this observation is an artefact of linear regression.

A comparison of the regression lines for T–P correlations (Figures 3c and 3d) for both groups shows that for the experimental group (slope = 0.70 and intercept = 1.05) the line is remarkably closer to the perfect positive regression line than it is for the control group (slope = 0.38 and intercept = 1.9). These observations indicate that the peers in the control group were overrating low ability students; a similar finding is noted by Falchikov (1995). This tendency seems to be present also with the experimental group but to a much lesser degree, thus bringing peer judgements closer to those of the teacher. The high *r* value (0.85) obtained for the experimental group strengthens this observation.

V Discussion

Judging from the relatively poor T–S correlations between the control and experimental groups, (*r*-values of 0.50 and 0.46, respectively), it

appears that peer feedback had not enabled the learners in this study to make judgements about their performances in a manner consistent with those made by the teacher. It can, therefore, be concluded that the empirical data presented in this study does not support Hypothesis 1 (see Section I).

The overall T–P correlations are considerably higher for the experimental group than the control group despite the fact that both groups had undergone the same level of training over the same period (see Section II.4 on Training). This observation lends strong support to Hypothesis 2 (see Section I), i.e., that peer feedback enhances learners' ability to make judgements on their peers' oral presentation skills comparable to those of the teacher.

1 Control group

For most second language students, speaking itself is a complex (and occasionally traumatic) task since it requires them to concentrate simultaneously on content, pronunciation, diction, eye contact, body language, etc. This complexity is unique to oral tasks compared to writing or listening tasks. With low ability students, this could be even more difficult. It is important to note here that the majority of the participants in the present study had obtained D or E grades in their Use of English Examination. Previous studies on self- and peer-assessments with EFL (English as a Foreign Language) students had found that it was difficult for them to make sound judgements of their own as well as their peers' speaking and learning abilities (Jafarpur, 1991). Thus, the low T–S or T–P correlations in the control group could be because of their language ability. It is, therefore, possible that the participants were unable to rate their performances realistically or in a manner comparable to those of the teacher. For instance, information gathered from the Student Evaluation Forms indicated that 58% of the students had said 'yes' and 42% 'no' when asked whether they had found it difficult to assess themselves. The reasons given were: 'I was nervous'; 'Forgot what I said'; 'cannot assess objectively'. When asked about rating their peers, 17% said 'yes', 75% said 'no' and 8% said 'sometimes'. The reasons given were that they had not done this before; could not differentiate as all of them were of the same level; and were sometimes unsure of their errors.

An interesting finding from the examination of the linear and non-linear regression exercises (Figures 3a to 3d) is that the data had exhibited an overwhelmingly linear regression relationship. This, together with a slope of less than 1 and an intercept larger than 0, lends credibility to the view that both peers and self were over-estimating the performances of low achievers, although to a slightly

smaller extent in the case of self-assessments. These findings are not unprecedented. Similar observations have been made by Mowl and Pain (1995) and Sullivan and Hall (1997). Orsmond *et al*. (1997: 363) suggested that students '[high achievers] are more self critical than they are judgmental [and low achievers] are less self critical, but more judgmental'. With regard to peer-assessment, Woolhouse (1999) observed that peers had difficulty in making 'honest' judgements. Oldfield and Macalpine (1995) noted that peers felt 'emotionally prejudiced' against giving low grades to their classmates. Kwan and Leung (1996) found that despite clear explanations and detailed discussions, there is no guarantee that students share the same understanding of the assessment criteria or marking scale as the tutor. The comment 'I know I am poor but cannot understand if I am 1, 2 or 3' supports such a claim (self- and peer-assessment evaluation forms).

Moreover, as Shore *et al*. (1992) suggested, judging dimensions such as appropriacy, fluency and clarity are very subjective. Some comments from the student evaluation forms were: 'could not judge our pronunciation'; 'could not judge confidence'; 'could not identify mistakes in grammar'.

2 *Experimental group*

Amongst all the correlation results and regression lines exhibited in Figures 3a to 3d, the *r*-value for T–P (= 0.85) for the experimental group is the largest and the corresponding regression line slope (0.70) the closest to 1 (the 'ideal' value). This suggests that peers can assess in a manner comparable to the teacher provided that they have had the benefit of peer feedback. Further, although the average ratings of the teacher and peers are different (2.89 and 3.08, respectively; Table 3), the *r*-value for T–P (= 0.85) is high. This shows that although peers have been more 'generous' than the teacher with their marks, more importantly PA has been more in line with the teacher than SA in its ability to discriminate good and poor performances.

With self-assessment, the *r*-value of 0.46 for T–S for the experimental group (Table 5) is the lowest amongst all the *r*-values exhibited (Figure 3b). This suggests that, in contrast to peers, the students have been unable to judge themselves in a manner similar to the teacher, despite having received the same levels of peer feedback and training as the peers. Could this be because of the lack of objectivity as Underhill (1987) had observed? In other words, were the subjective factors strong enough to mask the potential benefits derivable from peer feedback? Could the following comments recorded by the researcher during feedback sessions be interpreted as pointing to the

existence of strong subjective factors (greater subjectivity in speaking assessment as noted by Ross, 1998)?

While peers were saying (peer feedback sessions):

> 'You are much more confident this time compared to your last performance.'
> 'Your eye contact improved a lot.'
> 'You are much better prepared this time. You have given a lot of supporting detail.'

Self reflection included (self- and peer-assessment evaluation forms):

> 'Different people said different things, don't know who was right.'
> 'They cannot identify some of my mistakes and say I'm good.'
> 'Some of them [peers] are too subjective.'

If indeed there are strong subjective factors that need to be considered, what are the psychological factors leading to the subjectivity? The answers to these questions are beyond the scope of the present article. Clearly, future work is needed to resolve this issue.

Finally, it should be pointed out that the students in the study were not randomly sampled from the student population in the university (since they were D to E level students). It can be argued that the apparent lack of impact of peer feedback on the correlation between SA and TA may be an artefact resulting from the low proficiency or narrow range of the student population. Further research is needed to establish whether this was, indeed, an artefact. Until this is done, it would be difficult to generalize the conclusion of the lack of SA and TA correlation.

VI Conclusions

The present study has shown that when assessment criteria are clearly set (in this study, set by the teacher), peer feedback will enable students to make judgements of their peers comparable to those of the teacher. Considering that both the control and experimental groups consisted of students with similar English proficiency – all students got a D or E in Use of English, consisted of students from similar programs, had undergone the same level of training, and had followed the same procedure – it can be supposed that peer feedback had helped in achieving greater correlation between teacher- and peer-assessments. If this, indeed, is the case, then teacher assessment could be supplemented with peer-assessment at a lower cost in the context of oral skills. If peers can be involved in the task of assessment, teachers' time could be utilized more productively on issues related to improving their teaching techniques.

However, in view of the small number and narrow range of the participants, the results presented in this study need to be interpreted

with caution. As far as self-assessment is concerned, once again we need to bear in mind that the present study involves learners from a remedial English class with very little experience in being autonomous learners. The task of self-assessment (to a certain extent peer-assessment) is thus a novelty to them. Research studies involving peer- and self-assessment indicated that to enable students to perform these tasks effectively they need training and experience (Jafarpur, 1991; Adams and King, 1995; Freeman, 1995; Pond *et al*., 1995).

Further work needs to be carried out using a broader range of participants drawn from varying levels of ability. Further, to avoid the possibility of misinterpreting the questionnaire items, bilingual questionnaires should be used. In addition, as some of the work on self-assessment (Stefani, 1994; Kwan and Leung, 1996) has suggested, learners should be involved in drawing up the criteria so that they can develop a better understanding of the assessment criteria. Finally, for understanding the psychological factors involved in learners' tendencies to over- or under-estimate their performances, individual interviews may be held after the self- or peer-assessment procedures.

VII References

Adams, C. and **King, K.** 1995: Towards a framework for student self assessment. *Innovation in Education and Training International* 32, 336–43.

Armanet, C.M. and **Obese-jecty, K.** 1981: Towards student autonomy in learning of English as a second language at university level. *ELT Journal* 36, 24–28.

Bachman, L.F. and **Palmer, A.S.** 1989: The construct validation of self ratings of communicative language ability. *Language Testing* 6, 14–29.

Boud, D.J. and **Tyree, A.L.** 1979: Self and peer assessment in professional education: a preliminary study in law. *Journal of the Society of Public Teachers of Law* 15, 65–74.

Cameron, L. 1990: Adjusting the balance of power: initial self assessment in study skills for higher education – a case study. In Bell, C., editor, *Assessment and evaluation*. London: Kogan Page, 63–72.

Davidson, F. and **Henning, G.** 1985: A self-rating scale of English difficulty. *Language Testing* 2, 164–69.

Falchikov, N. 1995: Peer feedback marking: developing peer assessment. *Innovation in Education and Training International* 32, 175–87.

Freeman, M. 1995: Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education* 20, 289–99.

Heilenmann, K.L. 1990: Self assessment of second language ability: the role of response effects. *Language Testing* 7, 174–201.

Hong Kong Examinations Authority 1990: Comparability study between TOEFL and CE English Language (Syllabus B). Unpublished. Hong Kong.

Hughes, I.E. and **Large, B.J.** 1993: Staff and peer-group assessment of oral communication skills. *Studies in Higher Education* 18, 379–85.

Jafapur, A. 1991: Can naïve EFL learners estimate their own proficiency? *Evaluation and Research in Education* 5, 145–57.

Kwan, K. and **Leung, R.** 1996: Tutor versus peer group assessment of student performance in a stimulation training exercise. *Assessment and Evaluation in Higher Education* 21, 239–49.

Miller, L. and **Ng, R.** 1994: Peer assessment in oral language proficiency skills. *Perspectives: working papers in the Department of English*. City University of Hong Kong.

Mowl, G. and **Pain, R.** 1995: Using self and peer assessment to improve students' essay writing: a case study from Geography. *Innovation in Education and Training International* 32, 324–35.

Oldfield, K.A. and **Macalpine, M.K.** 1995: Peer and self assessment at tertiary level: an experimental report. *Assessment & Evaluation in Higher Education* 21, 239–50.

Orsmond, P., Merry, S. and **Reiling, K.** 1997: A study in self assessment: tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education* 22, 357–67.

Oskarsson, M. 1984: *Self assessment of foreign language skills: a survey of research and development work*. Strasbourg: Council for Cultural Co-operation.

Pond, K., Ul-Haq, R. and **Wade, W.** 1995: Peer review: a precursor to peer assessment. *Innovation in Education and Training International* 32, 314–23.

Rolfe, T. 1990: Self and peer-assessment in the ESL curriculum. In Brindley, G., editor, *Vol. 6: The second language curriculum in action*. Sydney: NCELTR, Macquarie University, 163–86.

Ross, S. 1998: Self assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing* 15, 1–20.

Shore, T.H., Shore, L.M. and **Thornton III, G.C.** 1992: Construct validity of self and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psycholology* 77, 42–54.

Stefani, L.A.J. 1994: Peer, self, and tutor assessment: relative reliabilities: *Studies in Higher Education* 19, 69–75.

Sullivan, K. and **Hall, C.** 1997: Introducing students to self-assessment. *Assessment and Evaluation in Higher Education* 22, 289–305.

Underhill, N. 1987: *Testing spoken language*. Cambridge: Cambridge University Press.

Wangsotorn, A. 1980: Self assessment in English skills by undergraduate and graduate students in Thai universities. In Read, J.A.S., editor, *Directions in language testing*. Selected Papers from the RELC Seminar on "Evaluation and Measurement of Language Competence and Performance". Singapore, April 1980. Singapore University Press, 226–40.

Williams, E. 1992: Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education* 17, 45–58.

Woolhouse, M. 1999: Peer assessment: the participants' perception of two activities on a further education teacher education course. *Journal of Further and Higher Education* 23, 211–19.

Appendix 1 Self-assessment questionnaire

Name: _____

Topic: _____

Date: _____

Rate yourself by using the scale:

Poor	**Unsatisfactory**	**Satisfactory**	**Good**	**Excellent**
1	2	3	4	5

A. Introduction	
1. Topic sentence – appropriate?	1 2 3 4 5
2. Topic sentence – interesting?	1 2 3 4 5
3. My opinion on the issue – clearly stated?	1 2 3 4 5
B. Body	
4. Details supporting the main points – sufficient?	1 2 3 4 5
5. Details supporting the main points – relevant?	1 2 3 4 5
C. Conclusion	
6. The main points – summarized?	1 2 3 4 5
D. Language Use	
7. Grammar – accurate?	1 2 3 4 5
8. Fluency	1 2 3 4 5
9. Pronunciation – words clearly pronounced?	
E. Manner	
10. Confidence (not nervous)	1 2 3 4 5
11. Confidence (depended very little on my notes)	1 2 3 4 5
12. Eye contact	1 2 3 4 5
F. Interaction	
13. Non-verbal interaction with the audience (facial expressions, gestures)	1 2 3 4 5
14. Verbal interaction (involving the audience during the talk by asking questions and encouraging them to respond)	1 2 3 4 5

Appendix 2 Peer feedback form

While your classmate is presenting the talk, note down your comments to be used in the feedback session.

Student name: _____

Date: _____

Introduction
Topic sentence –
Opinion on the issue –

Body
Details supporting main points –

Conclusion
Summary of main points –

Language Use
Grammar –
Fluency –
Pronunciation –

Manner
Confidence –
Eye contact –

Interaction
Overall interaction with the audience –

Appendix 3 Self- and peer-assessment evaluation form: control group

Answer the following by ✓ or ✗. Where necessary, provide an explanation.
By the end of the practice session:

1. Did you improve your speaking skills?
   If YES, what did you improve? (Tick wherever appropriate)
   Confidence in giving a talk – Pronunciation –
   Fluency –
2. Did you improve your ability in identifying strengths and weaknesses in your classmates' talks?

Peer Assessment

3. Did you find it difficult to assess your classmates' talks?
   If YES, why? _____

Self Assessment

4. Did you find it difficult to assess your own talk?
   If YES, why? _____

Overall Evaluation

5. Did you find the whole exercise of self assessment and peer assessment:
   Useful –
   Why? _____

   Interesting –
   Why? _____

   Motivating –
   Why? _____

   Boring –
   Why? _____

6. Have you done this kind of task before?

Appendix 4 Self- and peer-assessment evaluation form: experimental group

Answer the following by ✓ or ✗. Where necessary, provide an explanation.
By the end of the practice session:

1. Did you improve your speaking skills?
   If YES, what did you improve? (Tick wherever appropriate)
   Confidence in giving a talk – Pronunciation –
   Fluency –
2. Did you improve your ability in identifying strengths and weaknesses in your classmates' talks?

Peer feedback

3. Did you feel free to comment on your classmates' weaknesses?
   If NO, why not? _____

4. Did you agree with all the comments your classmates made on your talks?
   If NO, why not? _____

5. Could you appreciate your classmates' comments made on the weaknesses of your talk?
   If NO, why not? _____

6. How many times did your classmates make positive comments on your talk?
   Never Sometimes Most of the time

Peer Assessment

7. Did you find it difficult to assess your classmates' talks?
   If YES, why? _____

8. Did you make any changes to your judgement of your classmates' talks after listening to others comments on their talks?

Self Assessment

9. Did you find it difficult to assess your own talk?
   If YES, why? _____

10. Did you make changes to your judgement of your own talk after listening to your classmates' comments on your talk?

Overall Evaluation

5. Did you find the whole exercise of self assessment and peer assessment:
   Useful –
   Why? _____
   Interesting –
   Why? _____
   Motivating –
   Why? _____
   Boring –
   Why? _____

6. Have you done this kind of task before?