Self-assessment and Peer-assessment: A Comparative Study of Their Effect on Writing Performance and Rating Accuracy

Parviz Birjandi [*]
Professor of TEFL, Allame Tabatabaee University, Tehran, Iran
&
Masood Siyyari [†]
PhD Student, Allame Tabatabaee University, Tehran, Iran

Abstract
Self-assessment and peer-assessment are two means of realizing the goals of educational assessment and learner-centered education. Although there are many arguments in favor of their educational benefits, they have not become common practices in educational settings. This is mainly due to the fact that teachers do not trust the pedagogical values and the reliability of learners' self- and peer-assessment. With regard to these points, this study aimed at investigating the effect of doing self- and peer-assessments over time on the paragraph writing performance and the self- and peer-rating accuracy of a sample of Iranian English-major students. To do so, eleven paragraphs during eleven sessions were written and then self- or peer-rated by the students in two experimental groups. The findings indicated that self-and peer-assessment are indeed effective in improving not only the writing performance of the students but also their rating accuracy. After comparing the effects of self- and peer-assessment on the writing performance and the rating accuracy of the participants, peer-assessment, however, turned out to be more effective in improving the writing performance of the students than self-assessment. In addition, neither of the assessment methods outdid the other in improving the rating accuracy of the students.

Keywords: Educational Assessment; Self-Assessment; Peer-Assessment; Writing Performance; Rating Accuracy

[*] *E-mail address*: pbirjandi@yahoo.com
Correspondence address: Faculty of Persian Literature and Foreign Languages, AllamehTabataba'i University, Allameh Tabataba'i St., Tehran, Iran

[†] *E-mail address*: masoodsiyyari@yahoo.com

Introduction

By the advent of educational assessment in opposition to psychometric testing, placing testing at the service of learning became one of the major goals to pursue in education (Gipps, 1994; Brown, 1998; Lambert & Lines, 2000). Among several methods and techniques through which the goals of educational assessment could be accomplished, the alternative means of assessment are considered most effective. These alternative means include the use of checklists, videotapes, audiotapes, teacher observations, journals, logs, conferences, portfolio, self-assessment, and peer-assessment (McKay, 2006; Brown, 1998; Brown & Hudson, 1998, 2002).

According to Brown and Hudson (1998), the alternative means of assessment require the learners to perform, create, and produce in real-world contexts or simulations. Besides, the nature of these methods is nonintrusive and lets students be assessed on everyday class activities. The tasks used in these methods represent meaningful instructional activities which concentrate on both the process and the product of learning. Higher-level thinking and problem-solving skills are also the indispensable tools for carrying out the assessment tasks, and the teacher's feedback about the task performance sheds light on both the strengths and weaknesses of the learners. In addition, human judgment rather than machine judgment, as well as open disclosure of standards and rating criteria are encouraged.

Among the alternative means of assessment, self- and peer-assessment have attracted so much attention in recent years owing to growing emphasis on learner independence and autonomy (Sambell, McDowell, & Sambell, 2006). In addition, self- and peer-assessment have been viewed as having significant pedagogical values. According to Brown and Hudson (2002), self-assessment requires less time to conduct in classroom. Moreover, the students are very much involved in the process of assessment, and this by itself can lead to learner autonomy and higher motivation (Dickinson, 1987; Harris, 1997; Oscarson, 1989). Topping (2003) also emphasizes that self- and peer-assessment are cognitively demanding tasks which require and encourage intelligent self-questioning, post hoc reflection, learners' ownership and management of learning processes, sense of personal responsibility and accountability, self-efficacy, and meta-cognition.

Despite this much support for self- and peer-assessment, they are less than often practiced in educational settings especially in language teaching. This is probably due to the fact that the ability of the learners to assess themselves accurately and reliably is doubted. Studies on the reliability of self- and peer-assessment have also added to the uncertainty of teachers and administrators about the learners' ability to do self- and peer-assessment reliably since the findings of these studies are quite contradictory (Oscarson, 1989; Patri, 2002); however, it should be born in mind that most of the unreliability of self- and peer-assessment is due to the way they are carried out, and better prospects could be imagined for self- and peer-assessment by controlling the effect of the intervening variables that might distort the final results. In the subsequent section, some of these intervening variables and factors are enumerated, and some major related studies in the literature are reviewed.

Review of the Literature

The literature review of self- and peer-assessment reveals that some factors have been found to account for inaccuracy in self- and peer-assessment. For instance, Blanche (1988) has concluded from a comprehensive literature review that students' accuracy in self-assessment depends on the linguistic skills and the materials used in assessment. Moreover, more proficient learners tend to underestimate themselves in self-assessment. Some factors such as past academic records, career aspirations, peer group, or parental expectations, and lack of training in self-assessment could also affect the subjectivity of learners in self-assessment. In addition, Davidson and Henning (1985), Blanche (1988), Janssen-van Dieten (1989), and Heilenmann (1990) have found that the level of language proficiency has an impact on the accuracy of language learners' self-ratings.

Brown and Hudson (2002), however, assert that "some of these problems can be overcome if the descriptions that students are referring to in rating themselves are stated in terms of clear and correct linguistic situations and in terms of exact and precise behaviors that the students are to rate" (p. 84). Moreover, Oscarson (1989) maintains that training in self-assessment, and naturally peer-assessment, can indeed end in promising results as far as rating reliability is concerned.

In the literature, most studies on self- and peer-assessment have focused on the validity and educational values of these practices; however, contradictory results have been reported. Many different reviews of these results have been reported by many authors. Although every author has reviewed his or her own selection of

studies (e.g., Topping, 2003; Dochy & Segers, 1999), they imply that there is still a long way ahead to resolve these validity issues (Patri, 2002; Matsuno, 2009). In the following, it is tried to recap some works mainly in the field of language teaching, touching a specific dimension of self- and peer-assessment, which might concern the present study.

The reliability of learners' self- and peer-ratings is a major concern in education including language teaching and assessment. There is still a lot of doubt whether results from self- and peer-assessment could be used for important decision makings such as certification, pass/fail, or placement purposes. LeBlanc and Painchaud (1985), however, conducted a sequence of experiments which led to the use of self-assessment as a placement test. Their findings were based on the high correlations between two self-assessment questionnaires, one on the four basic skills and the other on the communicative ability to deal with a situation, and the results of a proficiency test. Ross (1998) also found significantly high correlation coefficients between 254 adult English learners' self-assessment test matching their course book content, a related achievement test, and teachers' assessment.

Patri (2002) conducted a study on comparing teacher-, peer-, and self-assessment of oral presentation skills of undergraduate students of ethnic Chinese background. After the students were familiarized with the assessment criteria through some training sessions, they were put into two groups, one group conducting self- and peer-assessment in the presence of peer-feedback, and the other group without any peer-feedback. By analyzing the data mainly through Pearson correlations, significantly more agreement was found between the teachers- and peer-assessment in the presence of peer-feedback than between teachers- and self-assessment in either the presence or absence of peer-feedback, or between the teachers- and peer-assessment in the absence of peer-feedback.

Saito and Fujita (2004) conducted an almost similar study to Patri's (2002) which involved written performance. They found a striking similarity between the peer- and teacher-ratings of essay quality, but no similarity was observed between teacher- and self-ratings, and between peer- and self-ratings. Moreover, the self-raters made a mixed extreme group of both the most lenient and most severe raters. Saito and Fujita (2004) justify their findings by arguing that

> Subjective points of view indubitably involve other psychological factors such as students' self-esteem, self-confidence, a cultural value of modesty, habits of overestimating self-ability and the like. In the present study, self assessment may have tapped more into those other psychological domains. (p. 48)

In another study, Cheng and Warren (2005) held an investigation into the attitudes of learners toward peer-assessment, the reliability, and probable educational benefits of peer-assessment of oral and written language proficiency in English language programs. By comparing (1) the students' attitudes towards assessing both the English language proficiency and other aspects of the performance of their peers, and (2) the teacher- and peer-assessments, they found that students had a less positive attitude towards assessing their peers' language proficiency, but they did not score their peers' language proficiency very differently from the other assessment criteria. They further asserted that two main reasons accounted for most of the students feeling unqualified to assess their peers' language proficiency. The first reason lied in the learners' uncertainty as to what constituted proficiency, and the second reason resulted from the learners' belief that their linguistic competence was insufficient for the task.

In a recent study, Matsuno (2009) emphasizes that traditional approaches to measurement, such as true-score approach, do not adequately take into account rater severity/leniency and assessment criterion difficulty level. With regard to these limitations, Matsuno (2009) employed Multifaceted Rash Model (MFRM) to compare self- and peer-assessment with teacher assessment in university writing classes. In this study, a sample of adult Japanese students used essay evaluation sheets based on the ESL composition profile by Jacobs et al. (1981) to practice self- and peer-assessment. MFRM analysis revealed that probably due to the Japanese culture for showing modesty, self-raters, especially those who were high achieving writers, were overly critical toward themselves. Peer-raters did not show much variance, they were lenient, internally consistent, and their rating patterns had no bearing on their own writing performance. However, peer-raters rated low-achieving writers leniently and high-achieving writers severely, as well as the fact that peer-raters produced fewer bias interactions than the self- and teacher-raters.

Before Matsuno, Davidson and Henning (1985) had also conducted a Rasch-based microscale analysis on the self-ratings of some ESL learners whose self-ratings were found to be reliable by classical methods of estimation while it was not the case when the data was analyzed through Rasch Model. In other words, lack of response validity was observed in the data, making Davidson and Henning assertively conclude that "little confidence should be placed in these particular student self-ratings" (p. 176).

The effect of self- and peer-assessment on learning and rating abilities is another dimension of self- and peer-assessment which has attracted the attention of researchers. For instance, Jafarpur and Yamini (1990) examined to what extent training could improve the self- and peer-ratings of thirty adult junior English majors at university. This study, which involved a pretest, treatment, and posttest, made use of three self-assessment and two peer-assessment questionnaires based on those of Oskarsson (1981) as well as the English Placement Test (EPT) (Corrigan et al., 1978), and a cloze as criterion measures. Multiple correlation analyses revealed that there was more overlap between the peer-ratings and the criterion measures than the self-ratings. To see if the level of proficiency had any relationship with self- and peer-assessment accuracy, the students were divided into three groups according to their pre-test scores from the EPT. No meaningful relationship was observed since the three groups were too small. For the null results of this study, Jafarpur and Yamini offer the type of questionnaires utilized and the insufficiency of the answers elicited by questionnaires as possible reasons. Moreover, Tarone and Yule (1989) are cited as arguing that

> Even if the learner is honest and capable of accurate self-analysis, the choice of response will inevitably reflect each individual's interpretation of what the statements entail. One learner may interpret the statement 'I can describe my home to him' as involving a brief description of the external appearance of a house, while another may think that a full description of the internal layout with all the furniture is also required. If the first learner answers 'Yes' and the second learner answers 'No', then the teacher has no insight, via this format, into what these learners are capable of. (p. 136)

Finally, Jafarpur and Yamini found that training did not improve the self-rating accuracy of the learners; however, they hypothesize that this finding was owing to the insufficiency of the training. They further bring support from Bandura (1977, 1982), Bandura and Schunk (1981), and Boekaerts (1991), to assert that a prerequisite for fairly estimating one's own performance is the ability to properly appraise the skills of others; however, the training in this study helped the learners just that much to make better appraisals of their peers only.

Since affective/attitudinal issues influence the result of any practice which involves human being, the field of assessment is also considered no exception in this regard. Therefore, affective/attitudinal issues in the practice of self-/peer-assessment are worth reviewing here. In a descriptive study, Mendonça and Johnson (1994) interviewed a group of 12 English learners from different fields of study, who had peer-reviewed their essays. The results revealed that all the students found the peer-review helpful indeed. The students believed that peer feedback helped them identify errors that they themselves could not find on their own. Moreover, the peer-review was considered a valuable opportunity when students found if somebody could understand their paper or not. Comparing one's paper with another's and learning something new as a result was felt to be a positive experience. The majority of the students also found the comments of peers from different fields of study useful because they could better pinpoint unclear parts in the essays; however, two students found this experience somewhat annoying.

Given the findings of the above-reviewed studies, drawing definite conclusions in terms of the nature of self- and peer-assessment should be done with caution; however, a list of worthy points gleaned from the above studies are enumerated as follows to conclude this section.

1. The design quality of self-/peer-assessment questionnaires can play an important role in determining the quality and validity of responses (LeBlanc & Painchaud, 1985; Jafarpur & Yamini, 1990; Ross, 1998).
2. The nature and content of what is going to be self-/peer-assessed, such as the kind of skill, can affect the results of the self-/peer-assessment (Jafarpur & Yamini, 1990).
3. Results of self-/peer-assessment can vary based on how language-proficient the learners are (Davidson & Henning, 1985; Blanche, 1988; Janssen-van Dieten, 1989; Heilenmann, 1990).

4. The users of the self-/peer-assessment questionnaires or scales need to be trained on how to use the instruments. Modeling by expert raters or teachers is one recommendation in particular (Jafarpur & Yamini, 1990; Saito & Fujita, 2004; Cheng & Warren, 2005).
5. Affective/attitudinal issues and psychological factors such as students' self-esteem, self-confidence, a cultural value of modesty, habits of overestimating self-ability and the like can affect the way self-/peer-assessment are practiced (Cheng & Warren, 2005; Saito & Fujita, 2004; Matsuno, 2009).
6. Relativity, self-flattery, and mismatch between the self-/peer-assessment items and criterion skills can distort the results of self- and peer-assessment (Ross, 1998).
7. Self-assessment and in particular peer-assessment need to be accompanied by constructive feedback from the teachers or peers to be most effective (Patri, 2002).

Statement of the Problem and Research Questions

High-stake decision making based on test results is not the only concern of teachers and administrators. What comes prior to testing is the issue of learning itself in educational assessment. One practice which has been considered to promote learning while assessing the language ability of the learners is educational assessment which could be realized through self-assessment and peer-assessment. Beside the many arguments for the advantages of self- and peer-assessment in the literature (e.g. Blanche, 1988; Oscarson, 1989), this study can provide further empirical evidence in terms of how successful self-assessment and peer-assessment are in fulfilling their promising objectives. To pursue this goal, the effect of self-assessment and peer-assessment over time on the writing performance was studied to see if self-assessment and peer-assessment would contribute to improvements in students' writing performance. Moreover, to find which assessment method would improve the writing performance more, comparisons were made between the effects of self-assessment and peer-assessment on writing performance.

In general, there is not much trust in the capability of learners to assess their own language ability and that of others (Oscarson, 1989; Patri, 2002). Inaccuracy exists in every measurement, especially in the field of human sciences; however, one cannot ignore the fact that the accuracy of rating can improve if there is enough training and practice. This issue holds true in the case of expert raters;

however, it does not mean that learners cannot be good raters if they are provided with enough training and practice. Even some empirical evidence from the literature supports this issue (Huttonen, 1986, as cited in Oscarson, 1989). Therefore, further empirical investigation of this issue can help teachers trust the learners more with their capability to rate their own language ability or that of their peers. With regard to this objective, the present study investigated if the practice of self- and peer-assessment of writing performance over time would improve the accuracy of learners' self- and peer-ratings. The rating accuracy improvements were also compared to see which assessment method would improve the rating accuracy more.

The research questions formulated with regard to the above-mentioned objectives and research problems are as follows.

1. Can student-writers' self assessment significantly improve the quality of their writing performance?
2. Can student-writers' peer assessment significantly improve the quality of their writing performance?
3. Can student-writers' peer assessment improve the quality of their writing performance more than student-writer's self-assessment?
4. Can student-writers' self-assessment significantly improve their rating accuracy?
5. Can student-writers' peer-assessment significantly improve their rating accuracy?
6. Can student-writers' peer assessment improve their rating accuracy more than student-writer's self-assessment?

Given the above research questions, the following null hypotheses were posed:

$H_0$1. Student-writers' self-assessment cannot significantly improve the quality of their writing performance.
$H_0$2. Student-writers' peer-assessment cannot significantly improve the quality of their writing performance.
$H_0$3. Student-writers' peer-assessment cannot improve the quality of their writing performance more than student-writer's self-assessment.
$H_0$4. Student-writers' self-assessment cannot significantly improve their rating accuracy.

$H_0$5. Student-writers' peer-assessment cannot significantly improve their rating accuracy.

$H_0$6. Student-writers' peer-assessment cannot improve their rating accuracy more than student-writer's self-assessment.

Method

Participants

The participants of this study consisted of 198 adult Iranian male and female students studying different English language majors at undergraduate level, including English literature, English translation, and English language teaching. The participants were from Allame Tabatabee University, the South-Tehran teacher training branch of Islamic Azad University, and Alborz Higher Education Institute. The needed data were collected from the participants while attending the Advanced Writing Course which is a two-credit 16-week course normally offered to the students in the third term of the bachelor's program. Since intact classes were used, the classes were arbitrarily assigned to treatment and control groups to have semi-randomized participants (Mackey & Gass, 2005). Table 1 shows how the participants were assigned to the treatment and control groups.

Table 1
Participants Assignment to Groups

University name	Peer-assessment group	Self-assessment group	Control group
Allame Tabatabee University	*n* = 33	*n* = 0	*n* = 29
Islamic Azad University	*n* = 0	*n* = 35	*n* = 31
Alborz Higher Education Institute	*n* = 35	*n* = 33	*n* = 0
Total	68	68	60

Instruments

Writing scale. The writing scale employed for scoring the paragraphs of the participants was the ESL composition profile by (Jacobs et al., 1981). Three raters used the scale to score the participants' paragraphs, and the participants of the treatment groups used the scale for the purpose of self- and peer-assessment. It should be noted that this scale was not used in its original form since it only includes scoring rubrics and brief descriptors for every writing component and key

word (i.e., content, organization, vocabulary, language use, and mechanics). Therefore, all the descriptors and the components of writing ability were fully explained and illustrated in a separate pamphlet for both the participants and raters. The participants' pamphlet differed to some extent from that of the raters' since the scale was translated into Persian for the participants, and the wording of the descriptors was simpler and less technical, accompanied by more examples. What was finally added to both pamphlets was a set of anchor scripts receiving the different scores for each writing component on the scale. These anchor scripts were actually sample paragraphs from students who had formerly taken the course, and the raters had rated with high inter-rater reliability.

Placement (proficiency) test. The participants' proficiency level was determined by means of the Oxford Placement Test (OPT). According to Allan (2004), the developer of the test, OPT has been calibrated against the proficiency levels based on the Common European Framework of Reference for Languages (CEF), the Cambridge ESOL Examinations, and other major international examinations such as TOEFL. The OPT calibrations have been based on direct and indirect data from multilingual populations of test takers and expert judgments. Each test is divided into two sections (Listening and Grammar), each of 100 items. These sections are also integrated with reading skills and vocabulary in context. Although many supporting explanations have been provided about the item facility values, discrimination indices, item and inter-test reliability, concurrent validity, and predictive validity of the test, the concurrent validity of the OPT was further established by calculating the Pearson correlation coefficient between the OPT scores and a retired paper-based TOEFL scores of 32 of the participants. Table 2 presents the correlation coefficients between OPT and TOEFL subskills and total scores, which are acceptable.

Table 2
Correlations between OPT & TOEFL Subskills and Total Scores

		TOEFL structure	TOEFL listening	TOEFL reading	TOEFL total
OPT grammar	*r*	.71(**)	.83(**)	.91(**)	.89(**)
	p	.00	.00	.00	.00
OPT listening	*r*	.72(**)	.87(**)	.92(**)	.91(**)
	p	.00	.00	.00	.00
OPT total	*r*	.72(**)	.86(**)	.92(**)	.90(**)
	p	.00	.00	.00	.00
	n	32	32	32	32

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

Procedure

Proficiency test administration. In the beginning of the study, the OPT was administered to all the participants to determine their proficiency scores. Table 3 provides the descriptive statistics for the participants' proficiency raw scores out of 200.

Table 3
Descriptive Statistics on Groups Proficiency Scores

	n	Min	Max	*M*	*SD*
Control group	60	82	177	133.88	26.30
Peer-assessment group	68	62	188	135.97	28.78
Self-assessment group	68	78	182	130.02	26.10

Since the level of English language proficiency was a relevant factor to the writing performance of the participants, the proficiency means of the groups were compared to see how different the groups were from one another. Since the Kolmogorov-Smirnov and Shapiro-Wilk tests showed the data was not normally distributed ($p < .05$), the nonparametric Kruskal-Wallis test was used to compare the proficiency means of the groups, which showed the groups were not significantly different; $H = 1.84$, $df = 2$, $p > .05$.

Rater training. In addition to one of the researchers of this study, two raters, who are experienced English language teachers at institute and university level holding Masters and Bachelors in TEFL, rated the writing performances of the participants. The rater training was conducted based on the procedures of Educational Testing Service elaborated by Weigle (2002). To check the initial interrater reliability of the raters, 30 paragraphs by the self-assessment group on the pretest were rated by the raters, and the interrater reliability was calculated via intraclass correlation (ICC), which turned out to be acceptable, that is .92.

*Self-/peer-assessment training***.** After the administration of the pretest, the writing course actually started with a two-hour session on the basics of paragraph writing such as topic, topic sentence, supporting sentences, coherence, cohesion, etc. Most of the instructions were based on Arnaudet and Barrett's *Paragraph Development (*1990). In the second session, the ESL composition profile accompanied by the related pamphlet containing the full descriptors, illustrations, and anchor scripts was introduced to the participants. The third session was also spent on the scale elaboration, and then sample paragraphs including the ones

written on the pretest were given to the students to be rated based on the scale and the anchor scripts. The students ratings were then compared with those of the raters, and the rating ambiguities were tried to be resolved.

Data collection. After the pretest, scale introduction, and paragraph rating practice by the participants, one method of paragraph development was introduced to the students every session. Having done the book exercises, the students were given two topics from their book, out of which one of them was supposed to be chosen for paragraph writing. In the peer-assessment group, the participants exchanged their paragraphs with those of their peers for peer-assessment; however, the participants of the self-assessment group rated their own paragraphs. This was done for nine sessions since there were as a whole nine paragraph development methods introduced to the students. After the ninth session, a posttest was also administered to check the improvement of the participants in writing performance and rating accuracy. Every session, the participants' paragraphs from the previous session were rated by the raters, and the necessary feedback was given to the students of all the three groups by their instructors. The feedback involved written and oral comments on those aspects of the students' paragraphs which either needed revision or deserved praise. Some sample paragraphs were sometimes read aloud by the students to be rated by both the instructors and students together in the class. Moreover, the peer-assessment group participants compared their own ratings with those of the raters every session. It is clear that scale introduction, rating practice sessions, self-assessment, and peer-assessment were absent in the control group, but the classes were tried to be as similar as possible as far as the writing instructions, practice, and feedback were concerned.

Data Analysis and Results

The numerical data for this study came from paragraph writing performance scores given by three raters. The final scores were the average of the three ratings rounded to the closest integer. To come up with the rating error scores, the difference between the self-/peer-ratings and the criterion scores (average of the three raters' ratings) were calculated.

As regards the statistical tests employed in this study, it should be noted that nonparametric tests, including Kruskal-Wallis test and Wilcoxon Signed Rank Test, were employed to compare the means since the data was not normally distributed based on Kolmogorov-Smirnov and Shapiro-Wilk tests results ($p < .05$).

Investigation of Research Questions

The first two questions of this study asked whether student-writers' self- or peer-assessment can significantly improve the quality of their writing performance. To start with, the mean scores across sessions for all the groups, including the pretest and posttest, were calculated. Figure 1 demonstrates how the groups' means have changed over the sessions.

Figure 1

Writing Performance Improvement Trends

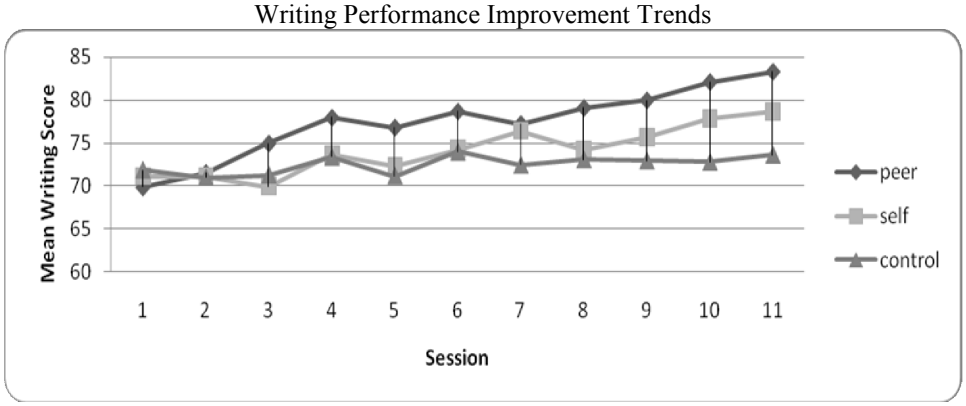


Figure 1 indicates that the three groups stared at close points with mean writing scores around 70 on the pretest. To see whether the groups were significantly different from one another as regards their writing performance on the pretest, their means were compared via Kruskal-Wallis test which showed that the groups were not significantly different; $H = 4.72$, $df = 2$, $p > .05$.

Although the writing performance trends for the three groups are upward, the peer-assessment group has ended in a higher mean writing performance (83.31) on the posttest in comparison to the self-assessment group's mean (78.68) and the control group's mean (73.63). Whether the average writing performance of the groups on the posttest was significantly better than that on the pretest, each groups' mean scores on the pretest and posttest were compared through Wilcoxon Signed Rank Test. For the self- and peer-assessment groups, the results of the test showed the differences were significant; control group $z = -3.36$, $p < .01$; peer-assessment group $z = -5.10$, $p < .01$; self-assessment group $z = -5.38$, $p < .01$. Therefore, null hypotheses 1 and 2 were rejected, and it could be asserted that student-writers' self-assessment or peer-assessment can significantly improve the quality of their

writing performance. However, the control group students, as seen above, had also improved significantly in their writing performance on the posttest in the absence of any self- or peer-assessment practice; $z = -3.36$, $p < .01$; therefore, it was necessary to take into account the mean writing performances of the groups on the pretest as a covariate, and then compare the writing performance of the three groups on the posttest through analysis of covariance (ANCOVA). This allowed for seeing whether the self- and peer-assessment groups improved significantly in their writing performance on the posttest in comparison to the control group. Moreover, research question 3, which asked whether student-writers' peer assessment can improve the quality of their writing performance more than student-writer's self-assessment, could be answered.

Although the data failed the normality tests, and the assumption of homogeneity of variances was not met based on Levene's test, F (2, 114) = 13.14, $p < .01$, the means of the groups were still compared through ANCOVA since ANCOVA, in the case of almost equal sizes ($n_1 = 36$, $n_2 = 39$), is believed to be robust enough even when these assumptions are not met (Leech, Barrett, & Morgan, 2005). Moreover, the most important assumption of ANCOVA, that is homogeneity of regression slopes was met too, Interaction F (2, 111) = 1.63, $p > .05$. Table 4 shows that the average writing performances of the groups on the posttest are significantly different taking into account the effect of the covariate; F (2, 111) = 50.35, $p < .01$.

Table 4
ANCOVA Results of Comparing Self- and Peer-assessment's Effects on Writing Performance

Source	*df*	*MS*	*F*	*p*	$p\eta^2$
Pretest mean (covariate)	1	748.35	40.76	.00	.26
Interaction	2	29.65	1.63	.20	.02
Posttest mean (dependent variable)	2	924.43	50.35	.00	.47
Error	113	18.35			

To find specifically where between the groups the mean differences existed, the Games-Howell test was run (Table 5). The reason for preferring this test to other post hoc tests was the fact that the groups had unequal variances.

Table 5
Games-Howell Multiple Comparison

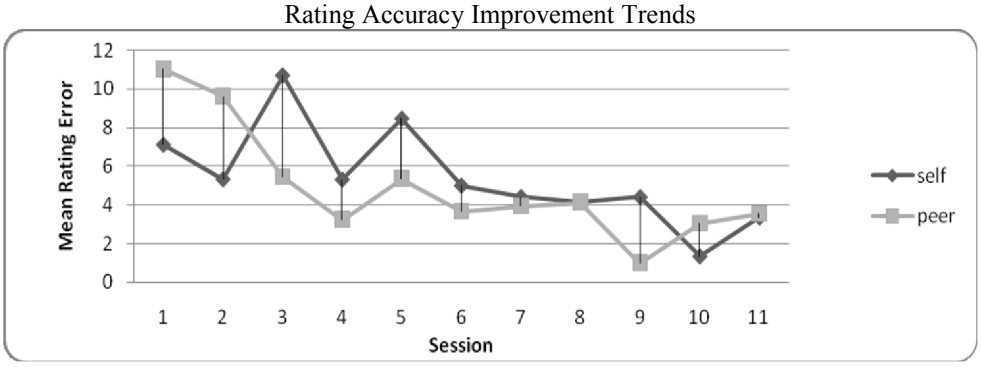
(I) CASE	(J) CASE	Mean Difference (I-J)	*SE*	*p*	95% Confidence Interval	
					Lower Bound	Upper Bound
peer	self	4.63*	1.26	.00	1.60	7.66
	control	9.67*	1.09	.00	7.04	12.31
self	peer	-4.63*	1.26	.00	-7.66	-1.60
	control	5.04*	.99	.00	2.65	7.44
control	peer	-9.67*	1.09	.00	-12.31	-7.04
	self	-5.04*	.99	.00	-7.44	-2.65

* The mean difference is significant at the .05 level.

As Table 5 shows, all the mean differences are significant ($p < .01$). In other words, the self- and peer-assessment groups improved in writing performance more significantly than did the control group. Moreover, the third null hypothesis was also rejected. That would mean that although both self- and peer-assessment improved student-writers' writing performance significantly, student-writers' peer-assessment can improve the quality of their writing performance even more than student-writer's self-assessment.

Research questions 4 and 5 asked whether student-writers' self- or peer-assessment can significantly improve their rating accuracy. To answer these questions, the mean rating errors across sessions, including the pretest and posttest, for the self- and peer-assessment groups were calculated. Figure 2 demonstrates how the means have changed over the sessions, and how differently the groups have performed in comparison to each other.

Figure 2
Rating Accuracy Improvement Trends



This graph indicates that the mean rating error of the self-assessment group on the pretest was 7.11, while it decreased to 3.34 on the posttest. For the peer-assessment group too, the mean rating error on the pretest (11.03) decreased to a much lower point on the posttest (3.53).

Whether or not the difference between the average rating errors of the groups was significantly lower than that in the beginning of the course, the mean rating errors of the pretest and posttest of the groups were compared through Wilcoxon Signed Rank Test. For both groups, the results of the test showed the differences were significant; peer-assessment group $z = -3.02$, $p < .01$; self-assessment group $z = -2.87$, $p < .01$. Therefore, null hypotheses 4 and 5 were rejected, and it could be claimed that student-writers' self- or peer-assessment can significantly improve their rating accuracy.

To answer the last question of this study, which asked whether student-writers' peer assessment can improve their rating accuracy more than student-writer's self-assessment, it was necessary to take into account the average rating errors of the groups on the pretest as a covariate, so that it would be determined which group had a lower average rating error on the posttest. As seen in Table 6, the homogeneity of regression slopes was met for ANCOVA; Interaction F (1, 66) = 3.91, $p > .05$; however, the data failed the normality tests, and the assumption of homogeneity of variances was not met based on Levene's test, F (1, 68) = 5.89, $p < .05$. Be that as it may, the means of the groups were still compared through ANCOVA for its robustness. Table 6 shows that the average rating errors of the

groups on the posttest are not significantly different taking into account the effect of the covariate; F (1, 66) = .003, p > .05. Thus, null hypothesis 6 was supported, implying that Student-writers' peer assessment cannot improve their rating accuracy more than student-writer's self-assessment.

Table 6
ANCOVA Results of Comparing Self- and Peer-assessment's Effects on Rating Accuracy

Source	*df*	*MS*	*F*	*p*	$p\eta^2$
Interaction	1	293.44	3.91	.05	.05
Pretest mean (covariate)	1	490.68	6.27	.01	.08
Posttest mean (dependent variable)	1	.25	.003	.95	.00
Error	67	78.19			

Discussion

The results of the data analysis above showed that all the three groups showed gradual improvement in their writing performance after nine sessions. However, comparison of the three groups' mean writing performance on the posttest showed that peer-assessment was the most effective practice in improving the writing performance of the participants, and the control group had the least degree of improvement in comparison to the other two groups. Support for the positive effect of peer-assessment has already been found in the literature by Cheng and Warren (2005). The reason for this finding is that self- and peer-assessment as two realizations of educational assessment come with many educational advantages such as the ones enumerated by (Gipps, 1994), namely having a formative nature, providing ongoing feedback, bringing about positive washback effect, promoting self-monitoring in learners, increasing self-esteem and motivation of learners and teachers, appraising via clear standards, and emphasizing mastery and progress. All being said, why peer-assessment turned out to be more effective than self-assessment must lie in the probable differences between peer-assessment and self-assessment. The findings of Mendonça and Johnson's study (1994) on the students' perceptions of peer-assessment can explain these differences well. Their study revealed that the students believed peer feedback helped them identify errors that they themselves could not find on their own. Moreover, the peer review was considered a valuable opportunity when students found if somebody could understand their paper or not; therefore, the students might be very competitively motivated to perform to impress their peers. Comparing one's paper with another's and learning something new as a result was also felt to be a positive experience by

the students. Evidently, peer-assessment might be beneficial to students in ways that might not be brought about by self-assessment since most of the above-mentioned points are dependent on the student's assessment of a peer's performance, which is absent in self-assessment. These points could be considered as reasons for peer-assessment to result in higher writing performance on the posttest in comparison to self-assessment.

In the present study, it was also found that the self- and peer-assessment groups showed improvement in their average rating accuracy after nine sessions of self- and peer-assessment practice, and the differences between these two groups in the degree they had improved in rating accuracy was not significant. In other words, the effects of the practice of peer-assessment and self-assessment on the rating accuracies of the two groups were almost similar. As Blanche (1988) and Oscarson (1989) have stated, training in self-assessment can increase the reliability of learners' self-ratings. In the present study, in addition to the first two sessions allocated to training the learners in assessing themselves and their peers, the learners had the opportunity to practice rating for nine more sessions. Thus, the more practice and training the students had, the more accurate they got in their ratings. This finding also agrees with the findings of LeBlanc and Painchaud (1985) which led them to the use of self-assessment as a placement test after they made sure students' self-ratings could be reliable enough. Similar supportive results on self- and peer-assessment reliability and validity have also been found by Ross (1998), Cheng and Warren (2005), Patri (2002), and Saito and Fujita (2004).

Contrary to the finding of the present study on the equal positive effect of self-assessment and peer-assessment on improving the rating accuracy of the participants, Jafarpur and Yamini (1990) found more agreement between students' peer-assessment and criterion measures than between students' self-assessment and criterion measures. They believe the reason for this finding is that for a learner to reach enough capability to appraise one's peer is much easier than to reach the rating ability necessary for self-appraisal; however, the present study showed that groups with almost equal training in peer- and self-rating showed equal improvement in rating accuracy after nine sessions. Of course, it should also be cautioned that the rating training in this study might have been that much enough to produce reliable results in self-assessment let alone peer-assessment. This is of course true if it is assumed that Jafarpur and Yamini's claim about peer-assessment ability as a prerequisite for fairly estimating one's own performance is correct.

Patri (2002) also found more agreement between the teachers-assessment and peer-assessment in the presence of peer-feedback than between teachers-assessment and self-assessment in either the presence or absence of peer-feedback, or between the teachers-assessment and peer-assessment in the absence of peer-feedback. The fact of the matter is that in the present study, there was no peer feedback, but only teacher's feedback to the students on their rating accuracy and writing quality. Now the question remains if the groups would have improved even more in rating accuracy if peer-feedback had been present. Probably further research, comparing the effect of teacher feedback and peer feedback, can resolve this question.

Saito and Fujita (2004) also found that peer-assessment of essay quality is more similar to instructor's rating than is self-assessment. To justify this finding, they put forward psychological factors such as students' self-esteem, self-confidence, a cultural value of modesty, and habits of overestimating self-ability as responsible for this finding. If this justification of theirs is assumed to be right, it seems that the present study was successful enough in controlling these intervening variables. It should be noted that during the present study, the students knew that no high-stake decision was to be made based on their peer-/self-ratings, and although the students showed no self-confidence and willingness in the beginning, the participants were always encouraged to know that they could make it as well as a teacher or expert rater. It should also be noted that the participants of this study were provided with a pamphlet containing full descriptions, illustrations, and anchor scripts on what constituted high quality paragraph writing performance. Moreover, the participants' rating improvement occurred in parallel with the improvement in their writing performance, thus the conclusion is that the more proficient the students get in their writing performance, the better they know what constitutes high quality writing performance, and naturally the more they show accuracy in their ratings. Knowing the exact criterion for acceptable performance is the factor that Cheng and Warren (2005) have also suggested in determining rating accuracy.

Conclusions

From the findings of this study, it could be concluded that both self- and peer-assessment can significantly improve the writing performance of learners in comparison to the common methods of teaching writing skill which might not give any opportunity to learners to assess their own performance or the ones of their peers. Therefore, language teachers, specifically those teaching the writing skill,

are highly recommended that they include more educational practices such as self- and peer-assessment in their teaching; this matter can guarantee both the learning of the students and increasing their motivation which is by itself an important factor in learning too. They also need not worry about the reliability of the students' self- and peer-ratings since learners can also get more and more accurate in rating as any expert rater does after enough training and practice is offered.

Received 5 February 2010
Accepted 7 March 2010

References

Allan, D. (2004). *Oxford placement test 1*. Oxford: Oxford University Press.

Arnaudet, M. L., & Barrett M. E. (1990). *Paragraph development: A guide for students of English (2nd ed.)*. New Jersey: Prentice Hall.

Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers, *RELC Journal,* 19(1), 75-96.

Brown, J. D. (Eds.) (1998). *New ways of classroom assessment*. Alexandria: TESOL Inc.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment, *TESOL Quarterly,* 32(4), 653-675.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency, Language *Testing,* 22(1), 93-121.

Davidson, F., & Henning, G. (1985). A self-rating scale of English proficiency: Rasch scalar analysis of items and rating categories, *Language Testing,* 2, 164-79.

Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge: Cambridge University Press.

Dochy, F., & Segers, M. (1999). The Use of Self-, Peer and Co-assessment in Higher Education: a review, *Studies in Higher Education,* 24(3), 331-350.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: The Falmer Press.

Harris, M. (1997). Self-assessment of language learning in formal settings, *ELT Journal,* 51(1), 12-20.

Heilenmann, K. L. (1990). Self-assessment of second language ability: The role of response effects, *Language Testing*, 7, 174-201.

Jacobs, H. J., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical Approach*. Massachusetts: Newbury House.

Jafarpur, A., & Yamini, M. (1995). Do Self-Assessment and Peer-Rating Improve with Training?, *RELC Journal,* 26(1), 63-85.

Janssen-van Dieten, A. (1989). The development of a test of Dutch as a second language: The validity of self-assessments by inexperienced subjects, *Language Testing,* 6, 30-46.

Lambert, D., & Lines, D. (2000). *Understanding assessment: Purposes, Perceptions, Practice*. London: Routledge Falmer.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument, *TESOL Quarterly,* 19(4), 673-687.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for Intermediate Statistics: Use and Interpretation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mackey, A., & Gass, S. M. (2005). *Second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms, *Language Testing*, 26(1), 75-100.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.

Mendonça, C. O., & Johnson, K. E. (1994*)*. Peer review negotiations: revision activities in ESL writing instruction, *TESOL Quarterly,* 28(4), 745–769.

Oscarson, M. (1989). Self-assessment of language proficiency: rationale and Implications, *Language Testing,* 6(1), 1-13.

Patri, M. (2002). The influence of peer feedback on self-and peer assessment of oral skills, *Language Testing,* 19(2), 109-131.

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

Saito, H., & Fujita, T. (2004*)*. Characteristics and user acceptance of peer rating in EFL writing classrooms, *Language Teaching Research,* 8(1), 31-54.

Sambell, K., McDowell, L., & Sambell, A. (2006). Supporting diverse students: developing learner autonomy via assessment. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 158-168). New York: Routledge,

Topping, K. (2003). Self and peer assessment in school and university: reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55-88). Dordrecht: Kluwer Academic Publishers.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.