Comparability of Holistic/Analytic Intra-reliability in Student/Teacher Assessment of Writing

MASOOD SIYYARI
Faculty of Foreign Languages and Literature
Islamic Azad University, Science and Research Branch, Tehran, Iran

Bio Data:
Masood Siyyari is PhD in TEFL from Allameh Tabataba'I University and a full time faculty member at Science and Research Branch of Islamic Azad University in Tehran, Iran. His main areas of interest are language testing/assessment and second language acquisition.

Abstract
Despite the many pedagogical benefits of self-/peer-assessment, they are not often practiced in the classroom, due to the fact that most teachers doubt learners' ability to do self-/peer-assessment accurately. Although several factors have been identified to affect self-/peer-assessment accuracy, the literature shows the rating accuracy of learners can improve if enough training is provided. Given the abovementioned supporting literature, it was hypothesized that learners, if provided with training and practice, may also have the potential to show behavior similar to that of expert-raters in terms of holistic and analytic intra-reliability. To test this hypothesis, having been trained to do self-/peer-assessment according to their group assignment, 136 English-major students conducted self-/peer-assessment of writing performance both holistically and analytically across 11 sessions. After correlating the students and raters' holistic and analytic scores and examining the variations among the correlations, it was found that students have indeed got the potential to show rating behaviors similar to those of expert raters and at times even show higher correlations. This paper closes with some implications these findings can have for theory and practice, and some new lines of research are recommended in the area investigated in this study.

Keywords: analytic scoring, holistic scoring, peer-assessment, self-assessment, teacher assessment, writing skill

Introduction

With the advent of educational assessment in opposition to psychometric testing, assessment in support 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 (Brown, 1998; Brown & Hudson, 1998; 2002; McKay, 2006).

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 the 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; Oscarson, 1989; Harris, 1997). 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 objectively is doubted by teachers (Oscarson, 1989). 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 (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 envisaged for self- and peer-assessment by controlling the effect of the intervening variables that might distort the final results.

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 concludes 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), Heilenmann (1990), and Jafarpur and Yamini (1995) 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.

With regard to the abovementioned supporting literature, it was hypothesized that learners, if provided with training and practice, may also have the potential to show rating behavior similar to that of expert-raters. Rating behavior can be defined in terms of the variance due to under/overestimation (i.e. strictness vs. leniency), variance in rating different skills, variance in rating different components of a skill, inter/intra-rater agreement or reliability, variance due to the choice of scoring method (holistic vs. analytic scoring), and holistic/analytic intra-reliability. This last instance of rating behavior is the kind of behavior on the side of raters and students which was compared in the present study. Put differently, this study investigated the extent of similarity between expert raters' holistic/analytic intra-reliability and that of students' in self- and peer-assessment of writing performance. To do so, steps described in the following sections of this study were taken.

Method

Participants

The participants of this study consisted of 136 Iranian male and female adult undergraduate students studying different English language majors, including English literature, English translation, and English language teaching, at Allameh Tabataba'i University, the South Tehran Teacher Training Branch of Islamic Azad University, and Alborz Higher Education Institute. The participants aged between 18 and 29, and the needed data for this study were collected from the participants attending the course Advanced Writing, 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 self- and peer-assessment groups by using a "semi-randomization procedure" (Mackey & Gass, 2005, p. 143). Table 1 shows how the participants were assigned to the groups.

Table 1
Participants Assignment to Groups

University	Peer-assessment group	Self-assessment group
Allameh Tabataba'i University	$n = 33$	$n = 0$
Islamic Azad University	$n = 0$	$n = 35$
Alborz Higher Education Institute	$n = 35$	$n = 33$
Total	68	68

Instrumentation

To provide the means for collecting the necessary data for this study, some instruments as follows were required.

Writing scale. The writing scale employed for scoring the paragraphs of the participants was the ESL composition profile by Jacobs et al. (1981) which has formerly been used for teacher-, self-, and peer-assessment by Saito and Fujita (2004) and Matsuno (2009) as well. Jacobs et al. (1981) have provided impressive indices on the reliability and validity of this scale, including interrater reliability coefficients between two, three, and four raters (ranging from .85 to .93), intercorrelations of the components of the scale (ranging between .64 and .89), Cronbach's coefficient alpha (.89), and results from a differential groups construct validity. For the purpose of the present study too, the same indices were computed on a sample of 30 paragraphs, the results of which are presented in Tables 2 and 3.

Table 2
Interrater Reliability Coefficients

	Number of raters	
	Two	Three
Reliability	.923	.929

Note. Reliabilities are calculated by intraclass correlation method

Table 3
Intercorrelations of the ESL Composition Profile Components

Component	Content	Vocabulary	Organization	Language use	Mechanics	Total
Content	1	.68**	.64**	.61**	.16	.87**
Vocabulary		1	.70**	.72**	.14	.84**
Organization			1	.54**	.30	.80**
Language use				1	.24	.83**
Mechanics					1	.41*
Total						1
Cronbach's alpha = .82						

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

* Correlation is significant at the 0.05 level (2-tailed).

As Table 3 demonstrates, almost all the coefficients are significant except the ones between Mechanics and the other components. This finding is of course justifiable on the grounds that as McNamara (1996) reasons, the component of Mechanics is concerned with surface editing of the text rather than the expressive aspect of writing; therefore, it was somehow expectable to find this components insignificantly correlated with the other components which are more linguistic and expressive by nature. To conduct a differential-groups construct validity study, the same participants' scores, which were collected in the beginning of a writing course, were compared with their own scores at the end of the writing course via paired-samples *t* test, which indicated a significant average increase from pretest to posttest; t (30) = -6.72, p <.01.

In the present study, the ESL composition profile was used for both raters' ratings and students' self- and peer-ratings. It should be noted that this scale was originally accompanied by scoring rubrics and brief descriptors for every key word and component to do with writing ability (i.e., content, organization, vocabulary, language use, and mechanics); however, all the descriptors and the components of writing ability were even further explained and illustrated by the researcher in a separate pamphlet for both the participants and raters. This explanatory pamphlet was mainly based on books by Jacobs et al. (1981), Hughey, Wormuth, Hartfiel, and Jacobs (1983), Kane (1988), and Arnaudet and Barrett (1990). The participants' pamphlet differed to some extent from that of the raters' since the scale was translated into Persian for the students, and the wording of the descriptors was simpler and less technical with more examples. Finally, a set of anchor scripts receiving the different band scores for each writing component on the scale was appended to both pamphlets. These anchor scripts were actually sample paragraphs from students who had formerly taken the course, and the raters had rated them with very high inter-rater reliability.

Proficiency test. Since language proficiency has been found to be a strong source of variation in rating accuracy of self- and peer-assessment (Davidson & Henning, 1985; Blanche, 1988; Janssen-van Dieten, 1989; Heilenmann, 1990), 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), and each section consists of 100 items. These sections are integrated with reading skills and vocabulary in context at the same time. Although a lot more has been said about this test in terms of its impressive item facility values, discrimination indices, item and inter-test reliability, concurrent validity, and predictive validity, the concurrent validity of the test was further established by calculating the Pearson correlation coefficient between 32 participants' scores on the OPT and a retired paper-based TOEFL. The Pearson correlation coefficients between the OPT and TOEFL subskills and total scores are presented in Table 4.

Table 4
Correlations between OPT and 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).

Data Collection Procedure

Proficiency test administration. In the beginning of the course, the OPT was administered to the students to determine their level of general English proficiency. Descriptive statistics on the groups' proficiency scores are presented in Table 5.

Table 5
Descriptive Statistics on Groups' Proficiency Scores

	n	Min	Max	*M*	*SD*
Peer-assessment group	68	62	188	135.97	28.78
Self-assessment group	68	78	182	130.02	26.10

Note. Maximum possible proficiency score = 200

The above proficiency raw scores were then calibrated against the OPT language level designation and the proficiency levels based on the Common European Framework (CEF), which indicated that their average language proficiency lay on the borderline between Lower Intermediate Modest User and Upper Intermediate Competent User (based on OPT), and B1 Threshold Independent User and B2 Vantage Independent User (based on CEF) (Allan, 2004). Since general English proficiency is an important factor in determining writing performance, it was ideal to have similar groups in terms of general English proficiency; therefore, the proficiency means of the groups were compared. Since the Kolmogorov-Smirnov and Shapiro-Wilk tests showed the data was not normally distributed ($p < .05$), the nonparametric Mann-Whitney *U* Test was used to compare the proficiency means of the groups, which showed the groups were not significantly different; $Z = -2.05$, $p > .05$.

Rater training. The researcher of this study as well as two EFL instructors, who were experienced English language teachers at institute and university levels and held Master's and Bachelor's degrees in TEFL, rated the writing performances of the participants. The rater training was conducted in several sessions by the researcher, who acted as the leader in the training process, based on the procedures of Educational Testing Service elaborated on by Weigle (2002) and the guidelines outlined by Jacobs et al. (1981). To check the holistic and analytic interrater reliability of the raters, 30 paragraphs by the self-assessment group on the pretest were rated by the raters, and the interrater reliabilities for holistic and analytic scorings were calculated via intraclass correlation (ICC), which turned out to be .94 and .92 respectively. It should be noted that the raters scored the paragraph first holistically and then analytically after all the holistic scorings of the paragraphs in one rating session were done. The holistic scores were given based on percentage. This was done based on the suggestion of Falchikov and Goldfinch (2000) in order to have both analytic and holistic scoring based on a similar range, and to provide the raters, and then the self- and peer-raters, with a familiar range.

Self/peer-assessment training and practice. 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, and cohesion. Most of the instructions were based on Arnaudet and Barrett's *Paragraph Development (*1990). The second session, the ESL composition profile accompanied by the related pamphlet containing the full descriptors, illustrations, and anchor scripts was introduced to the students. 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 first holistically and then analytically based on the scale and the anchor scripts. The students' ratings were then compared with those of the raters, and the rating ambiguities were discussed and resolved by the instructors during the session.

After the sessions spent on the introduction of the scale by the instructors and rating practice by the students, one method of paragraph development was introduced to the students every session. Having done the book exercises, the students were given a choice of two topics 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 afterwards since there were as a whole nine paragraph development methods introduced to the students. The students were told that self- and peer-rating data were to be used in partial determination of the class participation grade for each student.

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 both holistically and analytically, and the necessary feedback was given to the students. At times, some sample paragraphs were also read aloud by the students to be rated by both the teachers and students together in the class. Moreover, the participants in the peer-assessment group compared their own ratings with those of the raters every session.

Data Analysis

To test the hypothesis of this study, Pearson correlation coefficients of holistic and analytic self- and peer-ratings in every treatment session including the pretest and posttest were computed, the results of which are presented in Table 6.

Table 6
Correlation Coefficients of Holistic-Analytic Self- and Peer-ratings

Session	Holistic-analytic Pearson correlations	
	Self-assessment group	Peer-assessment group
Pretest	.90** n = 39	.90** n = 36
1	.87** n = 39	.84** n = 37
2	.79** n = 40	.91** n = 34
3	.71** n = 36	.84** n = 37
4	.85** n = 40	.84** n = 31

5	.80** n = 39	.78** n = 35
6	.85** n = 36	.89** n = 32
7	.93** n = 36	.87** n = 36
8	.91** n = 37	.84** n = 38
9	.88** n = 37	.84** n = 29
Posttest	.84** n = 41	.87** n = 39

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

To have a comparison of the holistic-analytic correlation coefficients between the two groups, the coefficients and the way they have changed across the sessions are illustrated in Figure 1.

Figure 1. Holistic-analytic correlation coefficients change trends over sessions

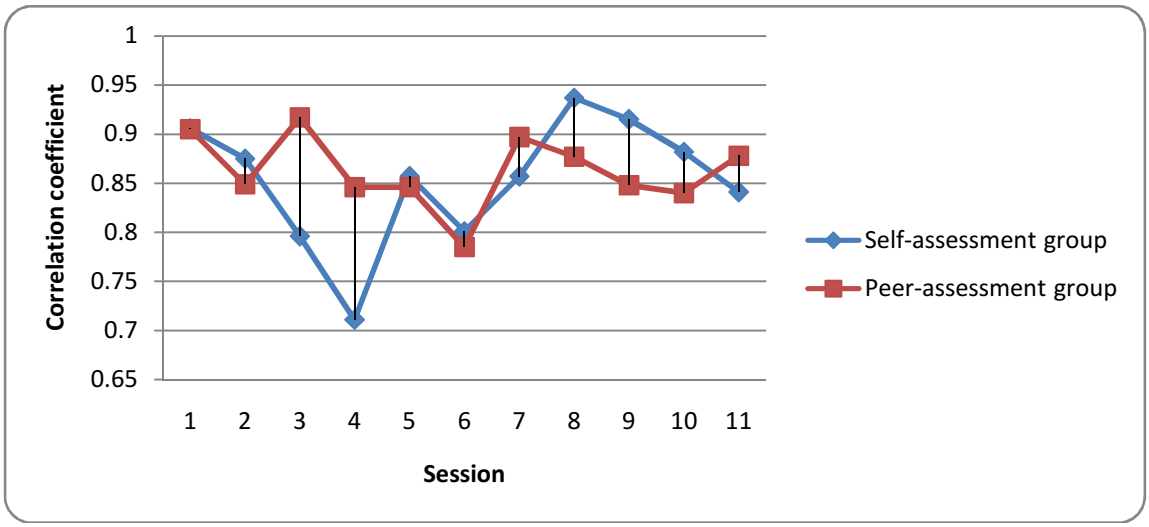


Table 6 and Figure 1 demonstrate that the correlation coefficients belonging to the self-assessment group range between .71 (treatment session 3) and .91 (treatment session 8). In the peer-assessment group, the correlation coefficients range between .78 (treatment session 5) and .91 (treatment session 2). All these correlations are significant (p < .01) with high enough common variances, although the fluctuations in correlation coefficients are evident over the sessions.

Although the above correlation coefficients are all significant with large effect sizes, comparing the holistic-analytic correlations coefficients of the participants with those of the raters could be interesting; therefore, all the holistic-analytic correlation coefficients of the raters over the sessions are provided in Table 7 and graphically shown in Figures 2 and 3 below to make the comparisons.

Table 7
Correlation Coefficients of Raters' Holistic-Analytic Scorings

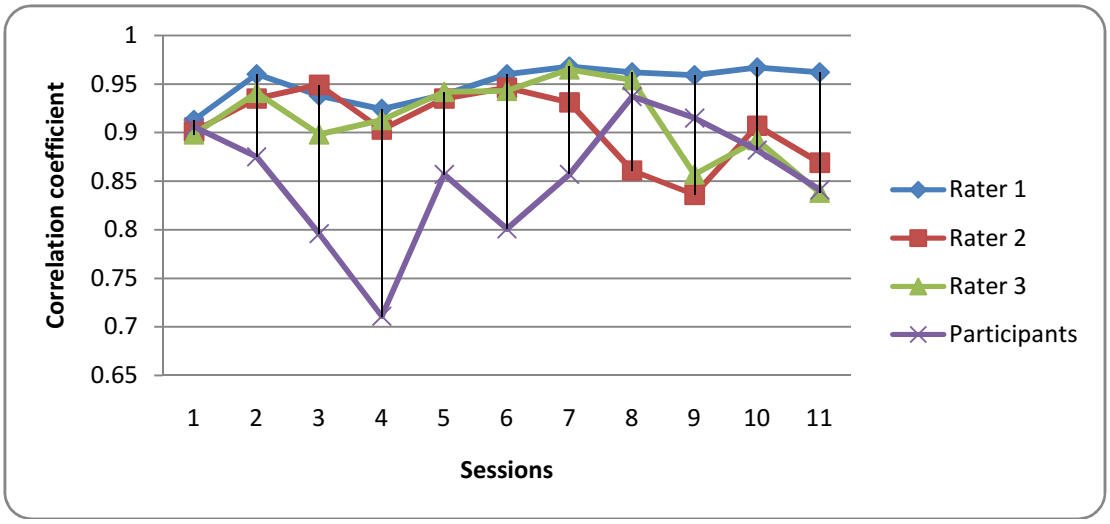
Session	Holistic/analytic Pearson correlations					
	Self-assessment group			Peer-assessment group		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
Pretest	.91** n = 65	.90** n = 65	.89** n = 65	.92** n = 64	.81** n = 64	.81** n = 64
1	.96** n = 63	.93** n = 63	.94** n = 63	.94** n = 61	.94** n = 61	.93** n = 61
2	.93** n = 61	.94** n = 61	.89** n = 61	.88** n = 59	.93** n = 59	.93** n = 59
3	.92** n = 49	.903** n = 49	.91** n = 49	.93** n = 54	.95** n = 54	.95** n = 54
4	.93** n = 60	.93** n = 60	.94** n = 60	.94** n = 57	.94** n = 57	.94** n = 57
5	.96** n = 57	.94** n = 57	.94** n = 57	.98** n = 60	.91** n = 60	.89** n = 60
6	.96** n = 52	.93** n = 52	.96** n = 52	.96** n = 50	.92** n = 50	.92** n = 50
7	.96** n = 50	.86** n = 52	.95** n = 52	.96** n = 63	.87** n = 63	.94** n = 63
8	.95** n = 53	.83** n = 53	.85** n = 53	.97** n = 57	.87** n = 57	.89** n = 57
9	.967** n = 52	.90** n = 52	.89** n = 52	.97** n = 48	.83** n = 48	.89** n = 48
Posttest	.96** n = 41	.86** n = 41	.83** n = 41	.96** n = 38	.85** n = 38	.82** n = 38

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

Considering the variations within each rater's ratings and between the raters' ratings implies that the participants' scorings were not that deviant from the raters' who were trained as experts to do the job.

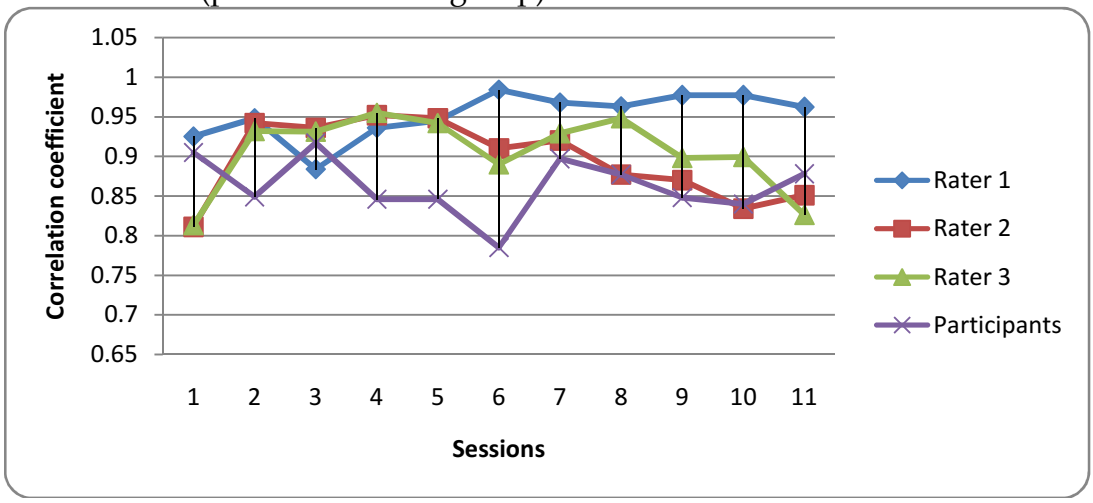
Moreover, the raters have also shown considerable fluctuations in their ratings over the sessions as the participants have. To be more precise in terms of this comparison, Figure 2 and 3 show the trends of correlation coefficient change over the sessions for both the raters and participants of both groups. Figure 2 shows that in the self-assessment group, the participants' correlation coefficients are noticeably deviant from the raters' only from session 2 to session 7; the rest of the sessions though show very close correlations between the raters and the participants; and interestingly the participants' correlation coefficients are at times even better than those of one or two raters' like sessions 8 and 9.

Figure 2. Raters and participants' holistic-analytic correlation coefficients change trends over sessions (self-assessment group)



For the peer-assessment group (Figure 3), the conditions is much better since the participants' correlation coefficients deviate from the raters' only in sessions 2, 4, 5, 6. In the rest of the sessions, the participants' correlation coefficients are clearly very close and sometimes even better than one or two of the raters', like sessions 1, 3, and 11. Finally, the noteworthy point is that the participants' correlation coefficients have shown gradual improvement and have got closer to those of the raters as the sessions have passed by.

Figure 3. Raters and participants' holistic-analytic correlation coefficients change trends over sessions (peer-assessment group)



Discussion and Conclusions

The data analysis above showed that there is a significantly high correlation between holistic and analytic scoring in self- and peer-assessment. Some insights were also gained by comparing the holistic-analytic correlation coefficients of the participants with those of the raters. The comparison demonstrated conspicuous variations

within each rater's ratings and between the raters' ratings, and the fact that the participants' scorings were not that deviant from the raters' who were trained as experts to do the job. Moreover, the raters showed considerable fluctuations in their ratings over the sessions as the participants did. To discuss the results in more detail, the trends of correlation coefficient change over the sessions for both the raters and participants of both groups were illustrated graphically, which showed that in the self-assessment group the participants' holistic-analytic correlation coefficients were noticeably deviant from the raters' only in a few sessions; sometimes they were very close, and in two sessions the participants' correlation coefficients were even better than those of one or two raters.

For the peer-assessment group, the conditions was even better since the participants' holistic-analytic correlation coefficients deviated from the raters' in fewer sessions than the self-assessment group participants. Like the self-assessment group, the participants' correlation coefficients were at times very close and sometimes even better than one or two of the raters'. Finally, the noteworthy point is that the participants' correlation coefficients showed gradual improvement and got closer to those of the raters as the sessions passed by. This is apparently due to the fact that the practice of self- and peer-assessment made the correlation coefficients improve. These findings are all significant since they show the rating behavior of learners as self- and peer-raters is very similar to that of expert raters. This similarity is also in agreement with several other findings and claims in the literature as mentioned before (e.g., Blanche, 1988; Oscarson, 1989; Ross, 1998; Patri, 2002; Brown & Hudson, 2002). The main implications of these findings are that learners' inaccuracy in conducting self- and peer-assessment can be a natural part of every rating process as it is the case for expert raters. Even expert raters might show inaccuracy and disagreement not only between each others' ratings but also within their own repeated and holistic-analytic ratings; however, this is the rating training and practice that can minimize these errors and disagreements between ratings. If this is the case, then why not providing the learners with the same rating training and practice, which can indeed result in improved rating as the results of this study revealed.

Suggestions for Further Research

Although the above findings indicate that learners as trained self-/peer-raters can at times show rating behavior like or even better than that of expert raters, further research needs to be conducted to show what variables and factors other than training quality might affect learners' rating accuracy and holistic-analytic scores agreement, and if so, how these factors can bring about the above influences. These factors can include issues to do with rating scales, training materials, number of sessions spent on training, and order of holistic and analytic rating.

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.). Englewood Cliffs, NJ: 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. (ed.) (1998). *New ways of classroom assessment*. Alexandria, VA: 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.

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

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

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*(2), 174-201.

Hughey, J. B., D. R. Wormuth, V. F. Hartfiel, & H. L. Jacobs. (1983). *Teaching ESL composition: Principles and techniques*. Rowley, MA: Newbury House.

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*(1), 30-46.

Kane, T. S. (1988). *Oxford essential guide to writing*. New York: Oxford University Press.

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

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

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.

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

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), 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.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht: Kluwer Academic Publishers.

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