Self-, Peer-, and Teacher-Assessments in Writing Improvement: A Study of Complexity, Accuracy, and Fluency

Hassan Soleimani[1] & Mahboubeh Rahmanian[2]

[1]Payame Noor University, arshia.soleimani@gmail.com

[2]Corresponding author, Payame Noor University, rahmanian74@yahoo.com

Received: 15/04/2014 *Accepted: 12/10/2014*

Abstract

Alternative assessment approaches and, among them, self-assessment and peer-assessment are becoming increasingly important in educational contexts. Designed to compare self-assessment, peer-assessment, and teacher-assessment, this study included 90 EFL students from 3 intact classes divided into 3 groups: self-assessment, peer-assessment, and teacher-assessment. After taking the TOEFL Proficiency Test (2004) and a writing pretest asking the participants to write a 150-word paragraph, the participants were trained upon the writing complexity, accuracy, and fluency (CAF) scale of Wolfe-Quintero et al. (1998). Before sitting the 2 posttests requiring the participants to write a 150-word paragraph, the self-assessment and peer-assessment groups assessed their own and peers' writings, respectively, whereas the third group had their teacher assess their writings. Results of one-way ANOVA demonstrated that teacher-assessment was not as effective as self-assessment and peer-assessment in terms of enhancing their writing proficiency. Results have important implications for educational organizations and curriculum designers who look for the most appropriate methods of teaching and testing.

***Keywords*:** Self-Assessment; Peer-Assessment; Teacher-Assessment; Complexity, Accuracy, Fluency (CAF); Writing Performance

1. Introduction

Since the rise of learner-centered approaches to learning, alternative assessment approaches have been considered as a substitute to the traditional psychometric assessment in educational contexts. Teacher-assessment as the long-standing psychometric assessment approach has been believed to be a valid and reliable method of measurement in pedagogical contexts (Huerta-Macias, 1995), whereas the validity and reliability of alternative assessment approaches were in doubt. Self-assessment and peer-assessment as two demonstrations of alternative assessment (Sambell, McDowell, & Sambell, 2006) are claimed to be the basic asset to independent learning (Oscarson, 1989) because they can help the participants to observe their own development metacognitively. Therefore, this approach might be

applied to motivate learners who are concerned about their own learning and educational goals (Harris, 1997).

Whereas many studies (e.g., Brown & Hudson, 2002; Topping, 2003) assert that an alternative approach to assessment is more effective, traditional assessment is yet the major concern of many teachers and organizations. Concerns on validity and reliability of assessment can be considered as the basic reason behind this decision. This study is determined to investigate whether these different approaches to assessment enjoy priority over one another in writing skill in terms of complexity, accuracy, and fluency (CAF).

2. Literature Review

A considerable amount of literature concerning the dichotomies of assessment approaches has left researchers/practitioners with many contradictories and unknowns. Traditional psychometric assessment, on the one hand, has brought about the satisfaction of teachers and administrators and, on the other hand, has not brought about motivation and independency of participants. Matsuno (2009) draws our attention towards teachers' temporal traits and assessment standards in order to propose a substitute to psychometric assessment. Understandably, he states that because these approaches and in particular the true-score do not consider seriousness or mildness of the teachers and measurement criterion, they need to be approached more cautiously.

Alternative assessment approaches and, among them, self-assessment and peer-assessment are becoming increasingly important in educational contexts because they can provoke more cognitive engagement on the part of the learners which in effect can bring conscious self-questioning, rethinking, and autonomous learning into the scene (Topping, 2003). Conversely, some variables like language proficiency (Heilenmann, 1990; Janssen-van Dieten, 1989), content of assessment scales (Ross, 1998), type of language skill to be assessed (Jafarpur & Yamini, 1995), psychological factors (Matsuno, 2009), and lack of scale training (Cheng & Warren, 2005) can threaten the validity and reliability of self-assessment and peer-assessment. Brown and Hudson (2002) hold 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" (p. 84). Validity and reliability of these approaches are in question because many studies (e.g., Dochy & Segers, 1999; Topping, 2003) have found contradictory results regarding this issue. Some studies have been undertaken to unravel whether employing self-assessment and peer-assessment are useful in different basic skills.

Painchaud (1985) conducted a series of trials in which they examined the correlation between self-assessment questionnaires on the four basic skills and the

communicative ability and a proficiency test. A high positive correlation was reported between the self-assessment of the four basic skills and the proficiency test. Similarly, Ross (1998) carried out a correlation study between self-assessment and teacher-assessment with 254 adult English participants in an achievement test and reported high positive correlation coefficients. Along the same lines, Patri (2002) carried out an investigation into the role of teacher-assessment, peer-assessment, and self-assessment in oral skills. The Chinese undergraduate students who were trained in applying the assessment scales were classified into two groups of self-assessment and peer-assessment with and without peer-feedback. Strong evidence of correlation was found with the teacher-assessment and peer-assessment with peer-feedback.

In another study, Saito and Fujita (2004) demonstrated that, in essay writing, there was a high correlation between peer-assessment and teacher-assessment, whereas teacher-assessment and self-assessment and peer-assessment and self-assessment revealed no such significant correlation. Saito and Fujita (2004, p.48) mention "students' self-esteem, self-confidence, a cultural value of modesty, habits of overestimating self-ability and the like" in order to explain unreliability of self-assessment. In 2009, Matsuno published a paper concerning writing skill in which he compared self-assessment and peer-assessment with teacher-assessment. After the adult Japanese students were instructed to apply the Jacobs, Zingraf, Wormuth, Hartfiel, and Hugheys's (1981) ESL essay measuring profile, they were told to assess their own and their peers' essay works based on it. Her study produced results which corroborate most of the findings of previous works; that is, self-assessment was not in line with teacher-assessment. Matsuno (2009) holds that this reason might be due to the Japanese culture of the participants. To put it simply, he claims that modesty made better writers underestimate themselves. Likewise, peer-assessment was similar to previous studies because it demonstrated more consistency and less prejudice in comparison with self-assessment and teacher-assessment.

Given the findings of some studies (e.g., Cheng & Warren, 2005; Saito &Fujita, 2004) regarding the positive effect of training on measurement scales on increasing the reliability of self-assessment and peer-assessment, Jafarpur and Yamini (1995), in a study on 30 adult university students, reached different conclusions, finding no development in self-assessment in the presence of beforehand training. However, they attributed this inconsistency to lack of enough training on measurement scales.

In the Iranian educational contexts, the writing skill is not paid so much attention to. Whereas many top Iranian and non-Iranian universities in higher education, nowadays, select their students based on their scores in proficiency tests including writing skill, writing is regarded as a subsidiary skill in Iran. This issue

exerts an influence on the way Iranian learners perceive writing. To put it simply, they can hardly deal with writing opportunities in their academic positions. It is tempting to suggest that we can change this trend if we look at the assessment of writing skill in another light, that is removing teachers from being the only authority and submitting this responsibility to the learners themselves in order to make more independent and motivated learners.

On the basis of the above studies, literature offers contradictory findings about which one of self-assessment, peer-assessment, or teacher-assessment can give rise to a more effective curriculum. Regarding the existing background not only in the broad context but also in Iran as an EFL context, more studies need to be undertaken on different types of assessment concerning basic skills and in particular writing in order to unravel which one can give better outcomes.

2.1 Complexity, Accuracy, and Fluency (CAF)

Proficiency is a commonly used notion in applied linguistics and more specifically in SLA, and yet it is a concept difficult to define precisely. Some researchers (e.g., Ellis, 2003, 2008; Skehan, 1998) draw our attention to the fundamental components of proficiency: CAF. Consequently, the CAF triad has turned out to be the main issue of applied linguistics studies in the literature. As such, Housen and Kuiken (2009) highlight the significance of this triad by stating that "CAF have been used both as performance descriptor for the oral and written assessment of language learners as well as indicator of learners' proficiency underlying their performance" (p. 461).

Complexity as a recent component among the CAF triad is equated with the extent to which the learners can create versatile and delicate language, whereas accuracy has come to refer to the extent to which the learners make their attempt to produce error-free language (Ellis, 2008). Fluency, on the other hand, embodies the two concepts of access fluidity and attention control where the former pertains to elaborating the lexical items to their underlying meanings and the latter pertains to partial attention to language production (Segalowitz, 2007).

Proficiency can be regarded as one of the mostly adhered apprehension of language learners upon which the three constructive components of fluency, accuracy, and complexity have received much attention. Since the advent of alternative methods of assessment, there has been more room for conducting experiments which can unravel how self-assessment, peer-assessment, and teacher-assessment can contribute the CAF triad.

Because the existing accounts fail to resolve the contradictions in assessment field, the following null hypotheses were formulated:

- H_{01}: Self-assessment in writing tasks has no statistically significant effect on improving the CAF of writing performance of Iranian EFL students.
- H_{02}: Peer-assessment in writing tasks has no statistically significant effect on improving the CAF of writing performance of Iranian EFL students.
- H_{03}: Teacher-assessment in writing tasks has no statistically significant effect on improving the CAF of writing performance of Iranian EFL students.
- H_{04}: Self-assessment, peer-assessment, and teacher-assessment in writing tasks have no statistically significant difference in improving the CAF of writing performance of Iranian EFL students.

3. Method

3.1 Participants

The sample consisted of 90 preuniversity Iranian EFL female students (N = 90), aged between 17-19, with a mean age of 18 years who studied English at a high school in Isfahan, Iran. The participants were three intact classes each containing 30 students who were classified into three groups respectively (i.e., self-assessment, peer-assessment, and teacher-assessment groups). The three groups (i.e., three intact classes) studied the same English materials at high school and had somehow similar average English scores. The TOEFL Proficiency Test (2004) and its writing section were administered in order to ensure the homogeneity of the sample population in terms of proficiency and writing.

3.2 Instruments

The following instruments were utilized in order to collect the data more precisely.

3.2.1 Writing scale in terms of CAF

L2 proficiency and performance have been thought to be measured by the multicomponents of CAF in many applied linguistics studies (Ellis, 2003, 2008; Skehan, 1998). Therefore, this triad has been regarded as valuable measurement variables for assessing not only the language proficiency, but also the oral and written skills (Housen & Kuiken, 2009).

3.2.1.1 Complexity

According to a definition provided by Ellis (2003), complexity is "[t]he extent to which the language produced in performing a task is elaborate and varied" (p. 340). Kuiken and Vedder (2007) hold that grammatical complexity is a complex concept to grasp which is why many morphosyntactic measures can be found in L2

literature. They also consider a positive correlation between proficiency level and number of clauses within a T-unit (i.e., any independent clause with dependent clauses, phrases, and words subsumed under), number of dependent clauses within a T-unit, and number of dependent clauses within the number of clauses. Likewise, Wolfe-Quintero, Inagaki, and Kim (1998) have argued that passives, articles, relative clauses, and complex nominals can demonstrate the proficiency level of the learners. More advanced learners can apply some more detailed morphosyntactic structures such as: temporal, deictic, logical subordinators, and the like. In the current study, we used a well-known and applicable definition of grammatical complexity, that is the mean number of clauses per T-unit (Larsen-Freeman & Strom, 1977; Wolfe-Quintero et al., 1998). In other words, because the concepts of T-unit and clause were assumed to be teachable, we decided on the applicability of this definition of complexity by Wolfe-Quintero et al. (1998).

3.2.1.2 Accuracy

For Ellis (2003), the term *accuracy* refers to performing a language task without any errors. In writing performance Wolfe-Quintero et al. (1998) assess accuracy as the number of T-units which are with no errors, the number of T-units within a T-unit which are with no errors, and the number of errors within a T-unit. Kuiken and Vedder (2007) believe that because we rarely are able to find T-units with no errors in the production of low proficiency learners, the first two methods of measuring accuracy in Wolfe-Quintero et al.'s(1998) definition might not be suitable in assessing the accuracy level of the low proficiency students. In the present study, we applied the third definition, that is, the number of errors within a T-unit in order to measure the accuracy of the participants' writing performance for the simple reason that the sample was low-intermediate students.

3.2.1.3 Fluency

Ellis (2003) has provided a good definition of fluency in L2 literature: "the extent to which the language produced in performing a task manifests pausing, hesitation or reformulation" (p. 342). The definition of fluency which we adopted in this study was the one proposed by Wolfe-Quintero et al. (1998) as the mean number of words within a T-unit.

3.2.2 TOEFL proficiency test

The TOEFL Proficiency Test (2004) and the writing section were administered to the participants in order to determine whether they were in the range of homogeneity needed in terms of proficiency and writing skill. Kolmogorov-Smirnov normality index revealed that the test scores were normally distributed for both the TOEFL Proficiency Test (p =.123, p < .05) and the writing section (p =.742, p < .05). As a result, one-way ANOVA test was used in which the

Levene's Test of equality of variances clarified homogeneous variances for the experimental groups. Thus, we could not have any significant differences across the groups considering both English proficiency ($p = .291$, $p < .05$) and writing skill ($p = .735$, $p < .05$).

3.2.3 Writing pretest and posttests

A writing pretest was carried out on the sample which asked them to write a paragraph of about 150 words in 20 min about their own life. The reason for selecting this topic was an attempt to make the participants demonstrate their best in writing skill. To put it simply, if a participant knew more vocabulary items and grammar points concerning a specific occurrence in his or her life, the participant was free to write about it. Then, their paragraphs were assessed by two external raters who considered the CAF writing measurement scale of Wolfe-Quintero et al. (1998) and Azar's (1985) error categorization for correcting compositions as singular or plural of words, form of words, word selection, tense of verbs, deletion or addition of words, order of words, not complete sentences, spelling, punctuation, capitalization, article, vague meaning, and vague and not distinct sentence.

By the same token, a posttest and a delayed posttest were given to the three groups two days and two weeks after the 8th session, respectively. The posttests, like the pretest, asked the participants to compose a paragraph of about 150 words in 20 min concerning their own life. Correspondingly, the CAF writing measurement scale of Wolfe-Quintero et al. (1998) and Azar's (1985) error categorization played a very important role in the assessment procedure of the posttests by the two external raters.

3.3 Data Collection

In an attempt to make the three groups homogeneous concerning both proficiency and writing skill, the TOEFL Proficiency Test containing the writing section was given to the participants. After administering the writing pretest, along with the purpose of our study, two external raters were trained regarding CAF scale of measuring writing skill by Wolfe-Quintero et al. (1998) They were, then, familiarized with error categories which were introduced by Azar's (1985) guide for correcting compositions as singular or plural of words, form of words, word selection, tense of verbs, deletion or addition of words, order of words, not complete sentences, spelling, punctuation, capitalization, article, vague meaning, and vague and not distinct sentence.

In order to test whether the external raters, who were experienced English language university instructors holding M.A. in TEFL, enjoyed interrater reliability, they were required to rate the pretest paragraphs of the three sample groups. Afterwards, the interrater reliability was computed as .92 using intraclass correlation

(ICC) that unraveled to be acceptable. The experiment done to reveal the effectiveness of self-, peer-, and teacher-assessments in writing improvement was done during 8 sessions. To enable the participants to write the paragraphs more efficiently, the fundamental points of paragraph writing were explained at the beginning of the first session using Arnaudet and Barrett's (1990) *Paragraph Development*. The participants were then familiarized with Azar's (1985) Guide for correcting compositions. Moreover, the CAF measuring scale of Wolfe-Quintero et al. (1998) was introduced to them. Afterwards, they were asked to assess the paragraphs based on the three dimensions of L2 writing proficiency, that is complexity (i.e., he mean number of clauses within a T-unit), accuracy (i.e., the mean number of errors within a T-unit), and fluency (i.e., the mean number of words within a T-unit). As the issue of clause was assumed to be difficult for the participants, they were first trained upon that.

Once the instructions were completed, some sample paragraphs were analyzed using both the fundamental points of paragraph writing and Azar's (1985) error categorization. In the end of the first session, the sample groups were asked to rate the pretest paragraphs randomly, the reason being that we could find out whether they have learned the instructions properly. Thanks to the interrater reliability, we could identify the vague points the participants had encountered during the process of rating and provided them with adequate descriptions to remove them. From the second session thereafter the participants were taught an academic writing issue in order to not only write a paragraph about, but also assess a paragraph based on the group to which they were belonged.

Taking Arnaudet and Barrett's *Paragraph Development* (1990) into account, for eight sessions we could train the participants an academic writing issue (e.g., comparison and contrast). Because preparing the preuniversity participants to write on the concerned academic writing issues was somehow difficult, we only taught them two to three key grammar and vocabulary items of each writing issue of *Paragraph Development* (1990). In addition, the tangible and teachable definitions of the CAF triad writing scale and Azar's (1985) error categorization were introduced to the participants at the beginning of the first session in order to let the participants correct the paragraphs. To enable the participants to write a paragraph about the considered writing issue, they were asked to do its writing exercises which contained two topics of writing as well of which the participants were expected to select one. Afterwards, the researchers explained the writing and assessment tasks of each group. Fortunately, because the preuniversity participants had only eight lessons in their text books each containing reading, vocabulary, and grammar sections, the researchers had enough time to conduct the experiment which was done during their normal class time. Moreover, thanks to their teachers who introduced

this research as an extra activity that could contribute to the participants' total English score, we could make the most of the participants' abilities.

3.3.1 Tasks

Thanks to Arnaudet and Barrett's *Paragraph Development* (1990), at the beginning of each session, an academic writing issue was explained to the three groups in order to contribute the participants to do the writing exercises at the end of each lesson. As a matter of fact, two writing topics of the exercises were selected out of which the participants should write about one. Afterwards, the self-assessment group was required to rate their own papers, peer-assessment group had to rate their peers' papers, and the teacher-assessment group had their papers rated by their teacher. The ratings needed to be based on the CAF measuring scale of Wolfe-Quintero et al. (1998), employing Azar's (1985) error categorization. The ratings of each group in each session were then correlated with the external raters' ratings in order to aware or inform the participants about their critical points in rating leading to better and more precise ratings in the next sessions. Two days and two weeks after the 8th session, a posttest and a delayed posttest were administered to the participants respectively. As with the pretest, the posttests required the participants to write a 150-word paragraph in 20 min about their own life. Likewise, the two external raters assessed the paragraphs based on the CAF writing measurement scale of Wolfe-Quintero et al. (1998) and Azar's (1985) error categorization to unravel if self-assessment, peer-assessment, and teacher-assessment can be beneficial in developing the CAF triad of the learners.

3.4 Data Analysis

K-S normality index was utilized to determine whether the posttest scores of the participants were normally distributed. Based on the normal distribution of the data for the first (p = .21), second (p = .12), and third (p = .24) experimental group sat the $p < .05$ level using the posttest scores repeated measure ANOVA was conducted. Similarly, the data for the fourth hypothesis were normal (p = .14) and in effect one-way ANOVA was run.

4. Results

4.1 Homogeneity Test Results

In order to be confident whether the three groups were homogeneous, prior to the study, the TOEFL Proficiency Test and its writing section were administered. K-S normality index revealed normally distributed test scores not only for the TOEFL Proficiency Test (p = .123, $p < .05$), but also for the writing section (p = .742, $p < .05$), leading to applying one-way ANOVA. The Levene's Test of equality of variances unraveled that the variances of scores across the three groups were

homogeneous. In other words, no significant difference was found between the three groups of the participants in terms of English proficiency (p = .291, p < .05) and writing skill (p = .735, p < .05) suggesting homogeneous variances.

4.2 Findings

4.2.1 Testing the first null hypothesis

Using the guidelines attributed to Wolfe-Quintero et al.'s (1998) and Azar's (1985) error categorization, the participants of the first group were able to assess their own paragraph writings with respect to the CAF writing scale. The descriptive statistics of the self-assessment group is presented in Table 1 which reports the influences that self-assessment could exert on the participants:

Table 1. *Descriptive Statistics of Self-Assessment Group in Terms of CAF Writing Scale*

	Group	*M*	*SD*	*N*
Pretest	Fluency	5.75	.32	30
	Accuracy	18.50	.65	30
	Complexity	.59	.04	30
	Total	8.28	7.57	90
Posttest	Fluency	11.00	.26	30
	Accuracy	2.45	1.31	30
	Complexity	1.35	.06	30
	Total	4.93	4.40	90
Delayed Posttest	Fluency	8.58	1.76	30
	Accuracy	11.03	3.02	30
	Complexity	1.04	.23	30
	Total	6.88	4.72	90

Comparisons between the three paired groups were made using one-way repeated measures ANOVA. There were significant differences between the three times of conducting the experiment (see Table 2) at the p < .05 level, Wilks' Lambda = .038, F(2, 86) = 1096.77, p <.001. Moreover, a large effect size was found (d = .96) with respect to Cohen (1988) criteria. The data were quite revealing that the first null hypothesis was rejected because a statistically significant difference was found between the mean scores of paragraph writings at three points of time indicating that the CAF writing scale can be developed if the participants assess their own writings. The results of the Tukey post-hoc test to locate the differences, shown in Table 3, indicated significant differences in all the comparisons.

Table 2. *Repeated Measure ANOVA for Self-Assessment*

Effect		Value	*F*	Hypothesis *df*	Error *df*	*Sig.*	Cohen's *d*
Self-Assessment	Pillai's Trace	.96	1096.77[a]	2.00	86.00	.00	.96
	Wilks' Lambda	.03	1096.77[a]	2.00	86.00	.00	.96
	Hotelling's Trace	25.50	1096.77[a]	2.00	86.00	.00	.96
	Roy's Largest Root	25.50	1096.77[a]	2.00	86.00	.00	.96

Note. Design: Intercept Within Subjects Design: Self-Assessment.

Table 3. *Tukey Post-Hoc for Self-Assessment*

(I) Self-Assessment		(J) Self-Assessment		Mean Difference (I-J)	*SE*	*Sig.*[a]	95% Confidence Interval for Difference[a]	
							Lower Bound	Upper Bound
Dimension 1	1	Dimension 2	2	3.34*	.07	.000	3.17	3.52
			3	1.39*	.21	.000	.87	1.92
	2	Dimension 2	1	-3.34*	.07	.000	-3.52	-3.17
			3	-1.95*	.22	.000	-2.50	-1.39
	3	Dimension 2	1	-1.39*	.21	.000	-1.92	-.87
			2	1.95*	.22	.000	1.39	2.50

Note. Based on estimated marginal means.
[a]Adjustment for multiple comparisons: Bonferroni
*$p < .05$ level.

4.2.2 Testing the second null hypothesis

Table 4 illustrates the descriptive statistics of the peer-assessment group who assessed their peers' writing papers. It can be seen from the data that peer-assessment could improve the participants' writing in terms of the CAF triad scale:

Table 4. *Descriptive Statistics of Peer-Assessment Group in Terms of CAF Writing Scale*

	Group	*M*	*SD*	*N*
Pretest	Fluency	5.75	.32	30
	Accuracy	18.50	.65	30
	Complexity	.59	.04	30
	Total	8.28	7.57	90
Posttest	Fluency	8.49	.26	30
	Accuracy	11.49	.76	30
	Complexity	1.00	.05	30
	Total	6.99	4.46	90
Delayed Posttest	Fluency	8.11	.98	30
	Accuracy	14.60	1.73	30
	Complexity	.94	.13	30
	Total	7.88	5.72	90

One-way repeated measures *ANOVA* was utilized in order to reveal whether peer-assessment could create a difference on the participants' writing skill. As seen in Table 5, the results were significant at the $p < .05$ level, Wilks' Lambda = .11, $F(2, 86) = 348.46$, $p < .001$ with a large effect size ($d = .80$) considering Cohen's (1988) criteria. The results could reject the second null hypothesis, as a statistically significant difference was found between the participants' writings in the three tests. Tukey post-hoc test was applied (see Table 6) in which all the comparisons turned out to be significant:

Table 5. *Repeated Measure ANOVA for Peer-Assessment*

Effect		Value	*F*	Hypothesis *df*	Error *df*	*Sig.*	Cohen's *d*
Peer-Assessment	Pillai's Trace	.89	348.46[a]	2.00	86.00	.000	.80
	Wilks' Lambda	.11	348.46[a]	2.00	86.00	.000	.80
	Hotelling's Trace	8.10	348.46[a]	2.00	86.00	.000	.80
	Roy's Largest Root	8.10	348.46[a]	2.00	86.00	.000	.80

Note. Design: Intercept Within Subjects Design: Peer-Assessment.
[a]Exact statistic.

Table 6. *Tukey Post-Hoc Test for Peer-Assessment*

(I) Peer-Assessment		(J) Peer-Assessment	Mean Difference (I-J)	*SE*	*Sig.*[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
Dimension 1	1	2	1.28*	.04	.000	1.16	1.40
		3	.39*	.10	.001	.13	.65
	2	1	-1.28*	.04	.000	-1.40	-1.16
		3	-.89*	.11	.000	-1.17	-.61
	3	1	-.39*	.10	.001	-.65	-.13
		2	.89*	.11	.000	.61	1.17

Note. Based on estimated marginal means.
[a]Adjustment for multiple comparisons: Bonferroni.
*$p < .05$ level.

4.2.3 Testing the third null hypothesis

The participants of the third group did not experience assessing; instead, they had the routine method assessment. Table 7 compares the writing development of the teacher-assessment group at different times with respect to the components of CAF writing criterion:

Table 7. *Descriptive Statistics of Teacher-Assessment Group in Terms of CAF Writing Scale*

	Group	*M*	*SD*	*N*
Pretest	Fluency	5.81	.47	30
	Accuracy	18.50	.65	30
	Complexity	.59	.04	30
	Total	8.30	7.57	90
Posttest	Fluency	5.50	.26	30
	Accuracy	14.00	1.03	30
	Complexity	.93	.07	30
	Total	6.81	5.47	90
Delayed Posttest	Fluency	7.33	.92	30
	Accuracy	13.66	2.77	30
	Complexity	.84	.12	30
	Total	7.28	5.52	90

The differences which teacher-assessment could have on the participants were examined using one-way repeated measures *ANOVA*. Statistically significant differences were observed between the pretest, the posttest, and the delayed posttest (see Table 8), Wilks' Lambda = 0.13, $F(2, 86) = 279.40$, $p < .001$, $p < 0.05$. A clear large effect size ($d = .96$) could be identified in this hypothesis referring to Cohen's (1988) criteria. Consequently, the third null hypothesis was rejected and teacher-assessment could create some changes upon the participants' writing skill.

As for pinpointing the location of differences, Tukey post-hoc test, as shown in Table 9, was run which detected all comparisons significant:

Table 8. *Repeated Measure ANOVA for Teacher-Assessment*

Effect		Value	*F*	Hypothesis *df*	Error *df*	*Sig.*	Cohen's *d*
Teacher-Assessment	Pillai's Trace	.86	279.40[a]	2.00	86.00	.000	.86
	Wilks' Lambda	.13	279.40[a]	2.00	86.00	.000	.86
	Hotelling's Trace	6.49	279.40[a]	2.00	86.00	.000	.86
	Roy's Largest Root	6.49	279.40[a]	2.00	86.00	.000	.86

Note. Design: Intercept Within Subjects Design: Teacher-Assessment.
[a]Exact statistic.

Table 9. *Tukey Post-Hoc for Teacher-Assessment*

(I) Teacher-Assessment		(J) Teacher-Assessment		Mean Difference (I-J)	*SE*	*Sig.*[a]	95% Confidence Interval for Difference[a]	
							Lower Bound	Upper Bound
Dimension 1	1	Dimension 2	2	1.48*	.06	.000	1.33	1.64
			3	1.02*	.16	.000	.62	1.42
	2	Dimension 2	1	-1.48*	.06	.000	-1.64	-1.33
			3	-.46*	.14	.007	-.83	-.10
	3	Dimension 2	1	-1.02*	.16	.000	-1.42	-.62
			2	.46*	.14	.007	.10	.83

Note. Based on estimated marginal means.
[a]Adjustment for multiple comparisons: Bonferroni.
*$p < .05$ level.

4.2.4 Testing the fourth null hypothesis

In order to compare the writing performance of the three groups with respect to the CAF triad writing criterion, one-way *ANOVA* was used. Table 10 provides the descriptive statistics for the immediate posttests of the self-assessment, peer-assessment, and teacher-assessment groups. What is interesting is that the self-assessment group obtained the highest means scores in comparison with the peer-assessment and teacher-assessment groups:

Table 10. *Descriptive Statistics of All Groups in Terms of CAF Writing Scale*

		N	*M*	*SD*
Self-Assessment	Fluency	30	11.00	.26
	Accuracy	30	2.45	1.31
	Complexity	30	1.35	.06
	Total	90	4.93	4.40
Peer-Assessment	Fluency	30	8.49	.26
	Accuracy	30	11.49	.76
	Complexity	30	1.00	.05
	Total	90	6.99	4.46
Teacher-Assessment	Fluency	30	5.50	.26
	Accuracy	30	14.00	1.03
	Complexity	30	.93	.07
	Total	90	6.81	5.47

To determine whether there were significant differences between the three groups, one-way ANOVA (see Table 11) was carried out, $F(2, 267) = 5.06$; $p =$

.007, $p < .05$, thereby the fourth null hypothesis was rejected. Interestingly, Tukey post-hoc test (see Table 12) unraveled each of the comparisons significant excluding the peer-assessment and teacher-assessment groups ($p = .966$) between which no statistically significant differences were reported:

Table 11. *ANOVA of All Groups of the Immediate Posttest*

	SS	*df*	*MS*	*F*	*Sig.*
Between Groups	234.09	2	117.04	5.06	.007
Within Groups	6168.90	267	23.10		
Total	6402.99	269			

Table 12. *Tukey Post-Hoc for All Groups in Terms of CAF Writing Scale*

	(J) Assessment	Mean Difference (I-J)	*SE*	*Sig.*	95% Confidence Interval	
					Lower Bound	Upper Bound
Self-Assessment	Peer-Assessment	-2.05*	.71	.012	-3.74	-.37
	Teacher-Assessment	-1.87*	.71	.025	-3.56	-.19
Peer-Assessment	Self-Assessment	2.05*	.71	.012	.37	3.74
	Teacher-Assessment	.17	.71	.966	-1.50	1.86
Teacher-Assessment	Self-Assessment	1.87*	.71	.025	.19	3.56
	Peer-Assessment	-.17	.71	.966	-1.86	1.50

*$p < .05$ level.

5. Discussion and Conclusion

This study compared the traditional and more updated methods of assessment not only to unravel their immediate and long-term achievement in terms of CAF measuring writing scale, but also to identify the one with the most encouraging results. The CAF measuring scale of Wolfe-Quintero et al. (1998) and Azar's (1985) error categorization were resorted to in order to probe the hypotheses more elaborately. Afterwards, two days and two weeks after the 8th session of assessment two writing posttests were administered to the participants.

The first research hypothesis investigated how self-assessment can affect the CAF triad scale. Complexity which was defined by Wolfe-Quintero et al. (1998) as the mean number of clauses per T-unit reported an increase in the immediate posttest; however, a significant decrease was observed from the immediate to the delayed posttests. One unanticipated finding was the behavior of accuracy (i.e., the number of errors within a T-unit) in the self-assessment group because it demonstrated the maximum level of accuracy only in the immediate posttest and a significant decline in the delayed posttest. Fluency as the mean number of words

within a T-unit increased sharply in the immediate posttest, but had a decline in the delayed posttest. Therefore, the first hypothesis suggested that the self-assessment group provided significantly better results in the immediate posttest rather than the delayed posttest. That is, the self-assessment group who was to assess their own paragraph writings benefited more from the initial learning because the differences which self-assessment could have upon the participants with respect to the CAF triad writing scale were not stable. Thus, it can be explained that although previous studies (e.g., LeBlanc & Painchaud, 1985; Ross, 1998) had supported self-assessment due to some merits like being formative, having perpetual reflection, bringing about valuable washback effect, reinforcing metacognitve skills, enhancing affective factors such as motivation and self-esteem, and highlighting achievement (Gipps, 1994), our study indicated its significance but only for initial learning.

The second hypothesis dealt with the effect of peer-assessment on the CAF triad in writing performance of the second group. Complexity, as the first criterion, presented a significant increase in the number of clauses which the peer-assessment group had utilized within a T-unit in the immediate posttest, whereas a significant decline was reported between the immediate and the delayed posttests. By contrast, accuracy indicated a significant decline in the number of errors used in a T-unit in the immediate posttest. But there again, the peer-assessment group could not come up with this drastic influence within two weeks after the treatment. Like complexity, fluency reported significantly much better results in the posttest compared to the delayed posttest. Therefore, it can be stated that, in contrast to the above hypothesis, the second hypothesis brought about a somehow long-standing increase in the writing proficiency of the participants with respect to the CAF writing scale.

The observed behavior of the peer-assessment group can be explained using the results obtained by Mendonça and Johnson (1994) who found the learners' presupposition that their peers are more capable of detecting their errors than themselves, can be enumerated as the main reason behind more efficiency in peer-assessment. Furthermore, they believed that knowing the fact that the students' writings would be assessed by their peers can make the learners write more understandably and effectively. To mention another explanation for their result, Mendonça and Johnson (1994) state that when the learners assess their peer's works, they unconsciously compare their own writings with their peers' which, in effect, can contribute them to learn new points. The above mentioned statements may explain why the peer-assessment group benefitted more from a steady CAF-triad improvement.

The third research hypothesis concerned with teacher-assessment in the CAF scale improvement. Even though the third group showed a clear trend of increment in using the number of clauses per T-unit (i.e., complexity) in the posttest,

there could rarely be seen a sign of considerable decrease within two weeks. Along the same lines, it was quite revealing that the participants whom their authority assessed their paragraph writings reported a significant error decrease in the posttest rather than the delayed posttest. Fluency, however, gave us interesting results because the teacher-assessment group postponed their improvement to two weeks after the treatment, something which could hardly be seen in the other two groups. Thus, it can be concluded that the findings of the third hypothesis were not in line with the previous studies. The CAF scale demonstrated different results for the teacher-assessment group that is even though teacher-assessment could not improve fluency in initial learning, accuracy and complexity corroborated the earlier findings. Surprisingly, passage of time allowed the teacher-assessment group to write more fluently but with less accuracy and complexity. On the whole, the results obtained by teacher-assessment group are not satisfying which can be attributed to the context (Teasdale & Leung, 2000) in which the students were assessed. In other words, an EFL context requires more motivation, self-esteem, reflection, and metacognition (Gipps, 1994) on the part of the learners which can hardly be attained by teacher-assessment. Although teacher-assessment is not as advantageous as self-assessment and peer-assessment with respect to the CAF triad, its steady results in accuracy and complexity do not let us ignore it completely. Along the same lines, Gipps (1994) believe that constant assessment of teachers which can give rise to "solid and broadly-based understanding of a pupil's attainment" (p. 123) might in effect lead to this claim that "teacher-assessment may be seen as having a high validity in relation to content and construct" (p. 124). Similarly, Teasdale and Leung (2000) after a long debate over psychometric and alternative assessments, came in to these interpretations too.

The fourth hypothesis dealt with the behavior of all the experimental groups after the 8th session, comparing the mean writing scores of the participants in the CAF triad. It is somewhat surprising that accuracy, among the other CAF triad writing scale, reported the most significant differences for self-assessment and the least significant differences for teacher-assessment. Likewise, the data from complexity and fluency revealed that the self-assessment and teacher-assessment groups presented the most and least satisfactory results, respectively. Consequently, the results of the fourth hypothesis showed that, contrary to expectations, peer-assessment could not provide the most satisfactory results in writing development in comparison with the self-assessment and teacher-assessment groups. To put it simply, our findings differed from most of the published studies concerning the superiority of peer-assessment (e.g., Cheng & Warren, 2005; Matsuno, 2009; Patri, 2002; Saito & Fujita, 2004), they accord with some investigations (e.g., LeBlanc & Painchaud, 1985; Ross, 1998) in the literature which support the importance of self-assessment. The priority of the self-assessment group over the other two groups

might be explained by the fact that self-assessment provides some conditions for the learners in order to think more deeply to the outcomes of their own learning and thereby bring them with more responsibility and independency over their learning (McNamara, 2001). However, the teacher-assessment group demonstrated the minimum level of the CAF triad among the other three groups. Matsuno (2009) refers to teachers' temporal characteristics to explain the contributing factor behind this behavior. Nevertheless, the newly developed approaches to assessment have proven to be more benefiting to the learner due to inspiring more motivation (Harris, 1997) and independency upon learners that can let them contemplate their own learning process (Oscarson, 1989). On the other hand, the peer-assessment and teacher-assessment groups were not significantly different from each other suggesting that the development of academic writing with respect to CAF scale can be approximately similar for both the peer-assessment and teacher-assessment groups. Notwithstanding the issue of training upon the writing scale has been a controversial subject in self-assessment (e.g., Cheng & Warren, 2005; Jafarpur & Yamini, 1990; Saito & Fujita, 2004), the results of this study indicate that training the writing measurement scale to the self-assessment group can have the most striking findings.

The present study was designed to determine the effect of alternative assessment (i.e., peer-assessment and teacher-assessment) approaches and traditional psychometric assessment on the improvement of the writing skill in pedagogical contexts. Returning to the hypotheses posed, it is now encouraging to compare these findings with that found by Brown and Hudson (2002) and Topping (2003). However, our findings draw a more delineated line between the two different dichotomies of alternative assessment. In other words, whereas self-assessment can temporarily have the most interesting results in writing improvement of the participants, peer-assessment can have a more steady writing development. Interestingly, teacher-assessment compared to the other two groups provided us with not so much encouraging results in the CAF triad writing scale. This finding has important implications for educational organizations and curriculum designers who look for the most appropriate methods of teaching and testing.

A number of caveats need to be noted regarding the present study. The most important limitation lies in the fact that CAF can be considered as some important issues in writing skill and there still remains many other areas to be worked on in writing field. Another weakness of this study is that Wolfe-Quintero et al.'s (1998) CAF writing formulation might not provide us with precise and delicate definitions. Likewise, the technical T-unit which was the fundamental issue in assessing the paragraph writings can be somehow a difficult concept to grasp for intermediate participants of our research.

References

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

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing*, *9*(3), 227-257.

Azar, B. S. (1985). *Guide for correcting compositions, fundamentals of English grammar*. Englewood Cliffs, NJ: Prentice-Hall.

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.

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.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford 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*, 174-201.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461-473.

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal*, *5*, 8-11.

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

Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M. Pilar & G. Mayo (Eds.), *Investigating tasks in formal language learning* (pp. 117-135). Clevedon: Multilingual Matters.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27*, 123-134.

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

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

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333-349.

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.

Segalowitz, N. (2007). Access fluidity, attention control, and the acquisition of fluency in a second language. *TESOL Quarterly, 41*, 181-86.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Teasdale, A., & Leung, C. (2000).Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing, 17*(2), 163-184.

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. 55-88). Dordrecht: Kluwer Academic Publishers.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu: National Foreign Language Resource Center.