International Journal of Research Studies in Language Learning
2015 October, Volume 4 Number 4, 15-27

The effects of using peer, self and teacher-assessment on Iranian EFL learners' writing ability at three levels of task complexity

Mosmery, Parisa ✉
Islamic Azad University of Damavand, Science and Research Branch, Iran (info@mosmeri.com)
Barzegar, Reza
Islamic Azad University of Damavand, Science and Research Branch, Iran (Barzegar72@yahoo.com)

ISSN: 2243-7754
Online ISSN: 2243-7762

OPEN ACCESS

Received: 3 October 2014 **Revised**: 19 January 2015 **Accepted**: 20 January 2015
Available Online: 20 February 2015 **DOI**: 10.5861/ijrsll.2015.928

Abstract

In the present study, the effects of using peer-, self-, and teacher-assessment on nurturing students' accuracy in writing skill were analyzed and furthermore, the researchers intended to perform the study at three levels of task complexity (simple, medium and complex), to find out whether being engaged in more complex tasks will help students improve accuracy in writing. This study was performed based on a framework for task design proposed by Robinson (2001, 2003, 2005, & 2007). To fulfill the purpose of this study, 81 (48 female and 33 male) upper-intermediate EFL learners of Iranmehr language institutes were chosen among a total number of 117 by means of a Nelson English Language Proficiency Test as the homogeneity test. Afterwards they were divided into three groups of Teacher Assessment (control group), Peer- and Self- Assessment (as experimental groups). After a pre-test, the students were exposed to some training for writing at three different levels of task complexity. Each student was assessed based on the group s/he was placed in using a rating scale in peer, and self-assessment groups. Then the post-test was administered and the gathered data was analyzed. The results indicated that although all three methods of assessment led to the students' progress, however, Self-Assessment method was the most effective method and Teacher and Peer were the second and third respectively. Furthermore, although a rise in the complexity level led to increment in results consistently, the increment in the second level was much higher than the third one which was contributable to less concentration in the third level.

Keywords: peer assessment; self-assessment; teacher assessment; task complexity

The effects of using peer, self and teacher-assessment on Iranian EFL learners' writing ability at three levels of task complexity

1. Introduction

Student involvement in assessment can be typically realized in terms of peer assessment or self- assessment. The feature common between these activities is that in both, certain criteria and standards are enjoyed by the students mostly applied to make judgments. In self-assessment, students judge their own work, while in peer assessment they judge the work of their peers (Falchikov & Goldfinch, 2000). Peer assessment plays a very significant role in formative assessment through making students judge the efforts of their co-learners. Furthermore, it is an assessment procedure that deals with the products of student learning, PA can also be considered as a process of learning by its own. Although the ideas regarding PA's positive role in classroom assessment are widely recognized, concepts concerning student perspectives of assessing and being assessed by peers are not quite well known (White, 2009).

Peer learning and assessment are quite effective in terms of developing students' critical thinking, communication, lifelong learning and collaborative skills (Nilson, 2003). Topping (1998) stated that not only peer assessment can increase the amount of feedback, but it can also promote higher level of thinking (Cheng & Warren, 2005; Nilson, 2003; Oliver & Omari, 1999; Orsmond & Merry, 1996; Sivan, 2000). The direct involvement in the learning process enhances students' sense of ownership, responsibility and students' motivation (Sivan, 2000). It promotes active and autonomous learners (Orsmond & Merry, 1996; Sivan, 2000, cited in Peng, 2008). Student self-assessment is a most important formative classroom assessment technique. To improve the quality of students' learning is a purpose of such technique. It can also lead to modifications when teaching strategies have not resulted into the required learning outcomes. A number of educators argue that students often find external assessment by instructors or supervisors unfair. Therefore, students will be more confident to give more accurate information about their progress in case of enjoying the chance to assess themselves (Angelo & Cross, 1993; cited in Baniabdelrahman, 2010).

Self-assessment is considered as a basis and one of the pillars of independent learning (learner autonomy). One of the most important elements of self-directed learning is that the students to be given enough opportunity to assess their own progress and thus to focus on their own learning. In general, it is believed that self-assessment is a key learning strategy for autonomous language learning, which enables students to monitor their progress and relate learning to individual needs (Harris, 1997). While, Dlaska and Krekeler (2008) strongly believe that one of the most important goals of self-assessment process is being provided with the authority of monitoring from which the learners benefit greatly. Furthermore, getting the learners familiar with standards against which they should appraise their performance is a criterion to get the accuracy of the self-assessment (Miller, 2003).

Similarly, Oscarsson (1989) gives six logical bases for self-assessment procedures. First, he stresses that self-assessment promotes learning. It gives learners training in evaluation, which has beneficial consequences for language learning. Second, it raises the awareness of both students and teachers of perceived levels of abilities. Through self-assessment, learners are encouraged to look at course content more carefully, and develop evaluative attitudes toward what and how they learn. Third, self-assessment is highly motivating with regard to goal-orientation. Learners gain knowledge of learning goals through reflection. Fourth, the involvement of learners in the assessment process results in the learner's broader perspective within the area of assessment. Fifth, by practicing self-assessment, students take part in their own evaluation, sharing the burden of assessment with their teacher. Finally, self-assessment may have long-term benefits, as one of the main aspects of autonomous language learning is the ability to assess the progress that is made.

Along with assessment techniques, task Complexity was another focus of this study. The term task has been such an important factor in the syllabus design, classroom teaching and learner assessment that it has influenced educational policy-making in ESL and EFL settings (Nunan, 2004). The most important role for a language task is to confront learners with certain language problems in completing the task (Long, 1985). Several frameworks have been proposed for task classification and design in SLA (Skehan & Foster, 2001; Robinson, 1995, 2005). A central issue in task-based language learning concerns the effect of task complexity on linguistic performance. Several studies have investigated the impact of task complexity on different aspects of linguistic performance at different levels of L2 proficiency (e.g., Robinson 1995, 2001; Skehan & Foster 1999; Rahimpour 1999, 2007; Gilabert 2007). However, Most of these studies have focused on oral proficiency.

According to Robinson (2001), "task complexity is the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner". A well-known model of task complexity was employed in this study, i.e. "the Triadic Componential Framework" known as "Cognition Hypothesis" proposed by Robinson (2001). Cognition hypothesis claims that if dimensions of cognitive task complexity belong to different attentional resource pools (e.g., memory and attention) increase in task complexity along the so-called resource-directing variables (e.g., +/- few elements, +/- Here and Now, +/- reasoning demand) lead to higher complexity and greater accuracy of learner's output. Robinson (2001a, 2001b, 2003, and 2005) argues that increasing task complexity with respect to resource-directing factors enhances complexity and accuracy but reduces fluency.

1.1 Study

The focus of this study is on one hand the assessment types (peer, self- and teacher assessment) and on the other hand, the engagement with more complex tasks and observing their effects (if any) on the improvement of Iranian Upper-intermediate EFL learners' writing ability. As the ultimate analyses made use of inferential statistics, the authors decided to proceed with research questions. Therefore, the following questions were proposed:

- Do peer-, self- and teacher assessments have any statistically significant impact on the improvement of Iranian Upper-Intermediate EFL learners' writing ability?
- If yes, which of the assessment types (peer/ self/ teacher assessment) is more effective on the improvement of Iranian Upper-Intermediate EFL learners' writing ability?
- Does Iranian Upper-Intermediate EFL learners' engagement with complex task and assessment of them (increased task complexity) improve their writing ability?

2. Method

As the nature of variables and the context of research need to identify the differences between three different variables (peer and self-assessment versus teacher assessment as control group) at three different levels of complexity, it is needed to make use of methods based on analyzing of differences between means of independent variables (ANOVA). But as the settings of this research needs to make use of a set of ANOVAs for pretest and another set for post-test results in different levels of complexity, using ANOVA to study the effects of these multiple factors has a complication. In fact, in a 3-way ANOVA with factors x, y and z, the ANOVA model includes terms for the main effects (x, y, z) and terms for interactions (xy, xz, yz, xyz). All terms require hypothesis tests. The proliferation of interaction terms increases the risk that some hypothesis test will produce a false positive by chance. This increases the chance for occurrence of type I error and falsely find a significant effect for factors which do not have a real effect on the dependent variable (due to repeating the same test a number of times). Testing one factor at a time hides interactions, but produces apparently inconsistent experimental results (Montgomery, 2001). So, it is probable to estimate every single equation at a time.

2.1 Participants

Due to the nature of this survey, two different groups; namely learners and teachers are needed and each group needs to fulfill the requirements of the research, which is explained in the upcoming section.

Learners - The participants of this study were 81 upper intermediate students of EFL studying in Iranmehr language institute. The participants included forty-eight female and thirty-three male students, ranging from nineteen to thirty-five years old. The participants were also homogenized by a Nelson English language test, at the very beginning of the research. After the homogeneity test, the participants whose scores were one standard deviation above and below the mean were selected for the purpose of this study (81 out of 117). Most of the participants of the study were from Tehran and there were some students from other cities of Iran. Therefore, the cultural background of the students was considered similar.

Teachers - Due to the specific regulations of Iranmehr institute, the researcher could not teach all of the classes herself. Therefore, six teachers of the institute accepted to cooperate with the researcher and run the treatments of the study in their classes. The treatments were run in nine General English classes.

2.2 Instrumentation

In order to have a homogeneous population and a uniform data and thus to be able to generalize the results, a homogeneity test was administered which included a 50-item Nelson English Language Proficiency Test (section 350 A) by Fowler and Coe (1976). This multiple-choice test included cloze passages, vocabulary, structure, and pronunciation items. For the peer-assessment and self-assessment groups, the researcher provided a checklist i.e. “Modified version of O’Malley and Valdez-Pierce's (1996) assessment form for writing” based on which they were instructed to do their assessment. The mentioned checklists were piloted for checking reliability in an Iranian context. Before finalizing the checklist to be used in the study, they were evaluated by 10 testing authorities. The next instrument used for this research was a pretest and posttest, which contained three writing tasks (including all three levels of complexity) to be written by the participants before and after the treatment. (Three writing tasks in pretest, three writing tasks in treatment and three writing tasks in the post test). The writing tasks were different in pretest and post-test, due to the fact that it was intended to prevent from students’ impact of background knowledge they have gained during the pretest and treatment.

2.3 Inter-rater Reliability Test of the Results

Researchers in various fields often need to evaluate the quality of a data collection method. These data recorded on a rating scale are based on the subjective judgment of the rater. Thus the generality of a set of ratings is always of concern. Generality is important in showing that the obtained ratings are not the idiosyncratic results of one person’s subjective judgment. Inter-rater reliability, inter-rater agreement, or concordance is the degree of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges. Generally Cohen’s kappa is used for estimation of reliability; which ranges from 0 to 1 and represents the proportion of agreement corrected by chance. Therefore, we performed inter-rater reliability analysis by making use of Kappa statistic for both pretest and posttest results obtained from teachers assessments and found a strong inter-rater reliability for both pretest (Kappa =0.64 ; p <0.01) and posttest results (Kappa =0.83 ; p <0.01).

2.4 Procedure

Before carrying out the writing tasks and conducting SA, PA and TA, the participants were trained on the notions of peer and self-assessment. The researcher met with the participants and described the phases of the research to them and thoroughly explained the background and purpose of the study, and also a hand out was given to them about how to write a three-paragraph essay. Then, the pretest was administered. Three writing tasks at three levels of complexity were given to the students; that is, one task in each session was given to them.

In the next step, the researcher explained the concepts on the rating scale and provided examples and demonstrations of how to use the rating scale and the students were given the checklists. Then the students provided with three subjects for writing at three levels of task complexity during the run of the research. The students were required to write a text of around 200 words for each task, and to do one task in each session. The time allocated for each task depended on the complexity of each task. Then the teacher, the students themselves or the peers using a rating scale, evaluated the three tasks. Finally, the post-test, which contained 3 tasks at 3 levels of complexity, was administered and the data was analyzed.

2.5 Evaluation criteria

The following criteria were used in the evaluation of writing tasks:

Punctuation - periods, commas & semicolon; question marks & exclamation marks; capital letters (proper names, beginning of sentences); quotation marks ("…")

Spelling - correct spelling

Sentences - S-V-O/S-V-C formation; subject/verb agreement; conjunctions (and, but, then, for, so, yet, or…); prepositions (on, in, into, out of, above, below, over, under…); articles (the, a, an)

Others - neat handwriting; good spacing, alignment & indentation

Overall writing (paragraphs) - main idea (clearly stated & supported by the following sentences); well argued (makes sense); well organized (introduction, body, conclusion); variety

2.6 Complexity of the tasks:

As mentioned earlier, three different tasks were defined in a way that they were increasingly more complex as follows:

First writing task - In this task which was considered to be simple, "few elements" and "no reasoning" were given a plus, because the learner is required to describe only one object and no reasoning is required because the task is a descriptive one and since the task requires a description of a person, event, or object in the past, a minus is given to the here-and-now variable. A plus is given to planning because the participants will be allowed to work in a group. Furthermore, a plus is given to the single task variable because it is just one task, the participants only describe the topic and are not required to do anything during the task. Finally, a plus is given to the prior knowledge variable because participants had once a task with similar topics. The time dedicated to this task was 40 minutes.

Table 1

Complexity in the First Task

Topic	Resource-directing	Resource-depleting
Describe a great teacher that you once had in your life	+ Few elements + No reasoning demands - Here & now	+ Planning + Single task + Prior knowledge
Describe a great apartment/house that you once lived in	+ Few elements + No reasoning demands - Here & now	+ Planning + Single task + Prior knowledge
Describe a great vacation that you once went on	+ Few elements + No reasoning demands - Here & now	+ Planning + Single task + Prior knowledge

Second writing task - The second task was considered to be more complex than the first task, because it required reasoning and also the topics were considered to be novel for the participants. The time dedicated for

completion of this task was 40 minutes.

Table 2

Complexity in the Second Task

Topic	Resource-directing	Resource-depleting
Motivate someone who has depression	+ Few elements - No reasoning demands + Here & now	+ Planning + Single task - Prior knowledge
Motivate someone (a fat one) to go on a diet	+ Few elements - No reasoning demands + Here & now	+ Planning + Single task - Prior knowledge
Motivate someone to learn English	+ Few elements - No reasoning demands + Here & now	+ Planning + Single task - Prior knowledge

Third writing task - The third task was considered to be more complex than the 2 other tasks, in that there was less writing time in task 3, i.e. 30 minutes, and the students weren't allowed to work in groups, Therefore, planning was given a minus.

Table 3

Complexity in the Third Task

Topic	Resource-directing	Resource-depleting
Discuss advantage & disadvantages of being single or getting married	+ Few elements - No reasoning demands + here & now	- Planning + Single task - Prior knowledge
Discuss which one is better: Money or Knowledge?	+ Few elements - No reasoning demands + Here & now	- Planning + Single task - Prior knowledge
Discuss advantages & disadvantages of playing computer games	+ Few elements - No reasoning demands + Here & now	- Planning + Single task - Prior knowledge

3. Results

The general purpose of multivariate analysis of variance (MANOVA) is to determine whether multiple levels of independent variables on their own or in combination with one another have an effect on the dependent variables. It is a generalized form of Univariate analysis of variance (ANOVA) and is used when there are two or more dependent variables. This helps to answer (Stevens, 2002):

- Do changes in the independent variable(s) have significant effects on the dependent variables?
- What are the interactions among the dependent variables?
- What are the interactions among the independent variables?

To do this, employed data needs to fulfill the following requirements:

- Assumption 1: The dependent variables must be distributed normally and their linear combination or any subset should also have a multivariate normal distribution. This is usually fulfilled via visualizing data and using histograms.
- Assumption 2: The population variance and covariance among the dependent variables should be the same across all levels of the factor. That is, variances for each dependent variable are approximately equal in all groups plus covariance between pairs of dependent variables is approximately equal for all groups. This is fulfilled by making use of Box's M statistic.

- Assumption 3: The participants are randomly sampled, and the score on a variable for any one participant is independent from the scores of this variable for all other participants. That is, each person's scores are independent of every other person's scores. This is commonly referred to as the assumption of independence. This requirement is already met by using Nelson's test; which was conducted earlier. However, MANOVA is robust to violations of multivariate normality and to violations of homogeneity of variance-covariance matrices if groups are of nearly equal size (N of the largest group is no more than 1.5 times the N of the smallest group). So, as these requirements will be fulfilled in the survey and the sample size for each group is the same, there is no need to worry about requirements and it is safe to use MANOVA for exploiting the underlying effects and relations between variables.

So, the research procedure could be summarized as: data would be first analyzed for normality, then covariance between data will be surveyed. After confirming the data requirements is met, the overall meaningfulness of model would be examined. Afterwards, being assured of equality of error term variances, between variable effects would be estimated. The next step would be devoted to provide pair wise comparisons and the final step is performing ad hoc tests to make sure no significant type I error is occurred. This process will be executed twice; once for pretest results and the second time for the post-test results. Then comparisons will be done and the major findings of the research in respect to the research questions will be presented.

3.1 Inter-rater Reliability Test on the results

The data recorded on a rating scale is based on the subjective judgment of the rater. So the generality of a set of ratings is always of concern. Generality is important in showing that the obtained ratings are not the idiosyncratic results of one person's subjective judgment. Inter-rater reliability, inter-rater agreement, or concordance is the degree of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges. It is an important measure in determining how well an implementation of some coding or measurement system works and quantifies the closeness of scores assigned by a pool of raters to the same study participants. Generally 70% of agreement is considered to be adequate. However, a better estimate of reliability can be obtained by using Cohen's kappa, which ranges from 0 to 1 and represents the proportion of agreement corrected by chance. This criterion is defined as:

$$K= (pa-pc)/(1-pc),$$

Where the pa is the proportion of times the raters agree and pc is the proportion of agreement we would expect by chance. Usually a 0.50 Cohen's Kappa is considered to be acceptable. Researchers usually take kappa values which at least 0.6 and most often higher than 0.7 as good level of agreement. However, the following table (Landis & Koch, 1977) can be used for the interpretation of the results:

Table 4

Kappa's results and interpretations

Kappa	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Refer to the table 5, the calculated Kappa (0.64; p <.0.01) shows that there is a substantial agreement between raters in the pretest and it is safe to use this data as a basis for statistical inference.

Table 5

Kappa values for pretest results

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Kappa Measure of Agreement	0.64	0.045	1.968	0.004
N of Valid Cases	30			

Note. a. Not assuming the null hypothesis. b. Using the asymptotic standard error assuming the null hypothesis.

Moreover, the results of the Kappa for posttest almost indicate a perfect agreement (Kappa = 0.83, p <.0.01) which is consistent with expectations that the more raters are got more familiar with the rating process, the more agreement happens to exist on the rating. So, Inter-rater reliability analysis using the Kappa statistic indicates a high level of agreement (concordance) between raters; which showed a substantial increment on posttest results.

Table 6

Kappa values for posttest results

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Kappa Measure of Agreement	0.827	0.028	0.219	0.006
N of Valid Cases	30			

Note. a. Not assuming the null hypothesis. b. Using the asymptotic standard error assuming the null hypothesis.

3.2 Pretest Results

The first step was checking for normality of the distribution of results, which showed that there were no meaningful difference between data distribution and normal distribution. MANOVA is most effective when dependent variables are moderately correlated (0.4–0.7). Fortunately correlations all are in the acceptable range and therefore we can easily proceed with MANOVA process.

Table 7

Pearson Correlations between teacher, self and peer groups

	Value	Teacher	Self	Peer
Teacher	Pearson Correlation	1	0.437**	-0.051
	Sig. (2-tailed)		0.000	0.634
	N	81	81	81
Self	Pearson Correlation	0.437**	1	0.099
	Sig. (2-tailed)	0.000		0.352
	N	81	81	81
Peer	Pearson Correlation	-0.051	0.099	1
	Sig. (2-tailed)	0.634	0.352	
	N	81	81	81

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

The next step is to run The Box's Test of Equality of Covariance Matrices in order to check the assumption of homogeneity of covariance across the groups using p < .001 as a criterion. As Box's M (7.315) is not significant (0.865) > (.001), so there are no significant differences between the covariance matrices. Therefore, the assumption is not violated and we can go further on estimating MANOVA.

Table 8

Box's Test of Equality of Covariance Matrice[a]

Box's M	7.315
F	0.574
df1	12
df2	14725.046
Sig.	0.865

Note. a. Design: Intercept + DIF.LEVEL

Afterwards, we verified the overall meaningfulness of model and as all statistics were meaningful at p < .05, the model as a whole considered to be meaningful. Then we conducted Levene's Test of Equality of Error Variances; which showed that the error variances could be considered to be equal and we can continue to specify the effects of complexity level on the assessments of three dependent variables (teacher, self and peer).

Table 9

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
Teacher	0.245	2	87	0.783
Self	0.176	2	87	0.839
Peer	1.055	2	87	0.653

Note. a. Design: Intercept + DIF.LEVEL

Then we explored the meaningfulness of effects of complexity levels on the evaluations of teacher, self and peer assessment; which showed that there was a statistical meaningful relationship between complexity level and assessments of teacher (p=0.28), self (p=0.41) and peer (p=0.31) which were all below 5 percent. So, the final estimates (MANOVA) for the effects of complexity on the evaluations of teacher, self and peer for the pretest is estimated as follows:

Table 10

Multiple Comparisons of different groups at different levels of complexity

Dependent Variable	(I) DIF.LEVEL	(J) DIF.LEVEL	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower	Upper
P.Teacher	LEVEL 1	LEVEL 2	-0.2921	0.18880	0.017	-0.7530	0.1688
		LEVEL 3	-0.3921	0.19981	0.015	-0.7589	0.2166
	LEVEL 2	LEVEL 1	0.2921	0.18880	0.017	-0.1688	0.7530
		LEVEL 3	-00.1000	0.16020	0.015	-0.3701	0.4120
	LEVEL 3	LEVEL 1	0.3921	0.19981	0.015	-0.2166	0.7589
		LEVEL 2	0.1000	0.16020	0.015	-0.4120	0.3701
P.Self	LEVEL 1	LEVEL 2	-0.3103	0.16547	0.002	-0.7143	0.0936
		LEVEL 3	-0.4489*	0.17512	0.036	-0.8764	-0.0214
	LEVEL 2	LEVEL 1	0.3103	0.16547	0.002	-0.0936	0.7143
		LEVEL 3	-0.1386	0.14040	0.020	-0.4813	0.2042
	LEVEL 3	LEVEL 1	0.4489*	0.17512	0.036	0.0214	0.8764
		LEVEL 2	0.1386	0.14040	0.020	-0.2042	0.4813
P.Peer	LEVEL 1	LEVEL 2	-0.0262	0.15340	0.000	-0.2935	0.4554
		LEVEL 3	-0.2927	0.16234	0.000	-0.6030	0.1896
	LEVEL 2	LEVEL 1	0.0262	0.15340	0.000	-0.4554	0.2935
		LEVEL 3	-0.0300	0.13016	0.000	-0.6054	0.0301
	LEVEL 3	LEVEL 1	0.2927	0.16234	0.000	-0.1896	0.6030
		LEVEL 2	0.0300	0.13016	0.000	-0.0301	0.6054

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

3.3 Posttest Results

As normality test of the distribution of results showed that there were no meaningful difference between data distribution and normal distribution, it was safe to go ahead with the process. Pearson correlations demonstrated that dependent variables were moderately correlated (0.39–0.55) and we could proceed with MANOVA. The result of Box's Test of Equality of Covariance Matrices was (0.416) > (.001), and demonstrated that were no significant differences between the covariance matrices. Moreover, overall model was meaningful and as all statistics were meaningful at p < .05. The Levene's Test of Equality of Error Variances showed that the error variances could be considered to be equal and we can continue to specify the effects of complexity level on the assessments of three dependent variables (teacher, self, and peer). Then the meaningfulness of effects of

complexity levels on the evaluations of teacher, self and peer assessment were explored and showed that there was a statistical meaningful relationship between complexity level and assessments of teacher (p=0.28), self (p=0.41) and peer (p=0.31) which were all below 5 percent. So, the final estimates (MANOVA) were made to answer the questions of research and the following inferences were made:

Table 11

Multiple Comparisons of different groups at different levels of complexity

Dependent Variable	(I) DIF.LEVEL	(J) DIF.LEVEL	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower	Upper
P.Teacher	LEVEL 1	LEVEL 2	-0.2921	0.18880	0.017	-0.7530	0.1688
		LEVEL 3	-0.3921	0.19981	0.015	-0.7589	0.2166
	LEVEL 2	LEVEL 1	0.2921	0.18880	0.017	-0.1688	0.7530
		LEVEL 3	-0.1000	0.16020	0.015	-0.3701	0.4120
	LEVEL 3	LEVEL 1	0.3921	0.19981	0.015	-0.2166	0.7589
		LEVEL 2	0.1000	0.16020	0.015	-0.4120	0.3701
P.Self	LEVEL 1	LEVEL 2	-0.3103	0.16547	0.002	-0.7143	0.0936
		LEVEL 3	-0.4489*	0.17512	0.036	-0.8764	-0.0214
	LEVEL 2	LEVEL 1	0.3103	0.16547	0.002	-0.0936	0.7143
		LEVEL 3	-0.1386	0.14040	0.020	-0.4813	0.2042
	LEVEL 3	LEVEL 1	0.4489*	0.17512	0.036	0.0214	0.8764
		LEVEL 2	0.1386	0.14040	0.020	-0.2042	0.4813
P.Peer	LEVEL 1	LEVEL 2	-0.0262	0.15340	0.000	-0.2935	0.4554
		LEVEL 3	-0.2927	0.16234	0.000	-0.6030	0.1896
	LEVEL 2	LEVEL 1	0.0262	0.15340	0.000	-0.4554	0.2935
		LEVEL 3	-0.0300	0.13016	0.000	-0.6054	0.0301
	LEVEL 3	LEVEL 1	0.2927	0.16234	0.000	-0.1896	0.6030
		LEVEL 2	0.0300	0.13016	0.000	-0.0301	0.6054

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

To explain the answers, it is needed to mention that the main concern of this research was to check if there is a statistically significance relationship between Self, Teacher and Peer Assessment and the writing ability of Iranian Upper-Intermediate EFL learners. The first finding is that there is such as relationship and as the power of model was high (around 78%) and all the used data fulfilled the statistical requirements, the reliability level is high and the reader can count on the results for any kind of decision making or further study. The next finding is that the Self-Assessment (with an overall improvement of 0.45 on a 0-5 scale) is the most effective way to improve writing ability of the Iranian Upper-Intermediate EFL learners. Afterwards, the Teacher-Assessment (with an overall improvement of 0.39 on a 0-5 scale) affects the writing ability and the Peer-Assessment (with an overall improvement of 0.29 on a 0-5 scale) is the least effective way to improve writing ability of the Iranian Upper-Intermediate EFL learners. However, the reader needs to be aware that all the three methods do have a positive effect on the writing ability. The last finding is that the increase in the complexity level constantly promotes the writing ability of Iranian Upper-Intermediate EFL learners. However, as the level of complexity rises, the increment pace decreases and the absolute increment in the second level (0.31 for Self, 0.29 for Teacher and 0.26 for Peer) is much more than the third level (0.14 for Self, 0.10 for Teacher and 0.03 for Peer).

4. Findings

As this research is organized in a way which enables the researcher to probe the statistically significance of the effectiveness of Self, Teacher and Peer Assessment on the writing ability of Iranian Upper-Intermediate EFL learners, the main finding to put is that Peer, self and teacher assessment have a statistically significant impact on the improvement of Iranian Upper-Intermediate EFL learners' writing ability. The researcher would like to mention that as the power of this model is high (around 78%) and all the used data fulfill the statistical requirements, the reliability level is so high and the reader can count on the results for any decision-making or

further study.

The next finding is that the Self-Assessment is the most effective way to improve writing ability of the Iranian Upper-Intermediate EFL learners. Afterwards, the Teacher Assessment affects the writing ability and the Peer Assessment is the least effective way to improve writing ability of the Iranian Upper-Intermediate EFL learners The last finding is that the increase in the complexity level promotes the writing ability of Iranian Upper-Intermediate EFL learners but as the level of complexity rises, the increment pace decreases, so the absolute increment in the second level is much more than the third level. The writer assumes that this is mainly contributable to the fact that as the level of complexity riser above a level, this increment leads to lack of concentration and subsequent poor performance in the test.

At level 1 of complexity, which is the easy level, the self-assessment group has outperformed compared to the other 2 groups, leaving behind the Teacher Assessment group at the second rank, with an increase of one and a half points and Peer Assessment group at the third level, with a half-point increase in the performance. Analysis on the second level of complexity has indicated that Self-Assessment group has occupied the first rank and the 2 other groups are mutually standing at the second level, both with a difference of 1 point between the pretest and post-test scores. Verifying the third level of complexity, i.e. the complexity tasks, on the other hand, has proved that Self-Assessment group has been standing on the first level of improvement, like the rest of complexity levels; Peer Assessment group has occupied the second rank, with a difference of 1 point between pretest and post-test and finally Teacher-Assessment group has fallen in the third rank, with a difference of half-points. Results of both overall analysis and item analysis have proved that Self-Assessment group has outperformed the two other groups in the writing performance. Peer-Assessment and Teacher Assessment groups have always stood at the second or third ranks, depending on the level of complexity.

4.1 Pedagogical Implications

Writing is considered as one of the most important skills that students of English as a foreign language (EFL) need to master. These students need to write for different purposes. In order to overcome the barriers students face while getting started to write and the fear of being evaluated in writing, it is recommended that teachers integrate new methods in their teaching writing process because of its considerable benefits. Integration of Peer- and Self-Assessment methods in writing instruction will let the students get some corrective feedback to correct their errors in writing, which gives them little anxiety compared to the teacher's feedback. This may help students to generate more mature writing compositions. This will also help students to promote the sense of independence in language learning and group work among their peers and overcome their fears in writing.

5. References

Baniabdelrahman, A. (2010). The effect of the use of self-assessment on EFL students' performance in reading comprehension in English. Retrieved from http://www.tesl-ej.org/wordpress/issues/volume14/ej54/ej54a2/2013

Cheng,W., & Warren, M. (2005). *Peer assessment of language proficiency*. SAGE Publications: Language Testing. http://dx.doi.org/10.1191/0265532205lt298oa

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 278-322. http://dx.doi.org/10.3102/00346543070003287

Fowler, W. S., & Coe, N. (1976). *Nelson English language tests*. London: Thomas Nelson and Sons Ltd.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education, 20*(3), 289-299. http://dx.doi.org/10.1080/0260293950200305

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics*, *45*(3), 215-240. http://dx.doi.org/10.1515/iral.2007.010

Harris, M. (1997). Self-assessment on language learning in formal setting. *ELT Journal, 51*(1), 12-14. http://dx.doi.org/10.1093/elt/51.1.12

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. http://dx.doi.org/10.2307/2529310

Long, M. H. (1985). A role for instruction in second language acquisition: task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modeling and assessing second language acquisition* (pp.77-99). Clevedon: Multilingual Matters.

Miller, C. R. (2003). Writing in a culture of simulation: Ethos online. In M. Nystrand & J. Duffy (Eds.), *Towards a rhetoric of everyday life: New directions in research on writing, text, and discourse* (pp. 58-83). Madison, WI: U. of Wisconsin Press.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New York: Wiley.

Nilson, L. B. (2003). Improving student peer feedback: *College Teaching*, *51*(1), 34–38. http://dx.doi.org/10.1080/87567550309596408

Nunan, D. (2004) *Task-based language teaching.* Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667336

O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* New York: Addison-Wesley.

Orsmond, P., & Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education*, *21*(3), 239–250. http://dx.doi.org/10.1080/0260293960210304

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: tutor and students' perceptions of performance criteria. *Assessment and Evaluation in Higher Education, 22*(4), 357-369. http://dx.doi.org/10.1080/0260293970220401

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

Peng, J. (2008) *Peer assessment in an EFL context: Attitudes and correlations.* Retrieved from http://www.lingref.com/cpp/slrf/2008/paper2387.pdf

Rahimpour, M. (1999). Task complexity and variation in interlanguage. In N. O. Jungheim & P. Robinson (Eds.), *Pragmatic and pedagogy: Proceeding of the 3rd Pacific second language research forum* (pp.115-134). Tokyo, Japan: Pac LRF.

Rahimpour, M. (2007). Task complexity and variation in L2 learners' oral discourse. *Working Papers in Language and Linguistics, University of Queensland,* 1-9.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524780.012

Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57. http://dx.doi.org/10.1093/applin/22.1.27

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies*, *21*(2), 45-105.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*(1), 1-33. http://dx.doi.org/10.1515/iral.2005.43.1.1

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. D. P.Garcia-Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7-26). Clevedon, UK: Multilingual Matters.

Robinson, P., Ting, S. C-C., & Urwin, J. (1995). Investigating second language task complexity. *RELC Journal*, *25*, 62-79. http://dx.doi.org/10.1177/003368829502600204

Sivan, A. (2000). The implementation of peer assessment: An action research approach. *Assessment in Education: Principles, Policy & Practice*, *7*(2), 193–213. http://dx.doi.org/10.1080/713613328

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*, 93-120. http://dx.doi.org/10.1111/1467-9922.00071

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.183-205). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524780.009

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276. http://dx.doi.org/10.3102/00346543068003249

White, E. (2009). Student perspectives of peer assessment for learning in a public speaking course. *Asian EFL Journal.* Retrieved from http://www.asian-efl-journal.com/pta_January_09.pdf