Akdeniz Language Studies Conference 2012

The effect of assessment type (self vs. peer) on Iranian university EFL students’ course achievement

Zainab Abolfazli Khonbi\textsuperscript{a}*, Karim Sadeghi\textsuperscript{b}

\textsuperscript{a,b}Department of English Language, Faculty of Literature and Humanities, Urmia University, Urmia165, Iran

Abstract

For education to be improved, there must be some indications of the efficiency in its developmental stages toward the ultimate objectives. In educational settings, the outcomes have traditionally been evaluated by authorities, i.e. by teachers themselves. Nowadays, parallel with the developments in teaching whereby students are seen as having more control over their own learning, the task of evaluation too is assigned to the students. In the present research, the effect of self- and peer-assessment was studied on Iranian university EFL students’ course achievement. The participants (19 and 21 students in self- and peer-assessment groups, respectively, and all from Urmia University) were pretested on their current Teaching Methods knowledge (based on Techniques and Principles in Language Teaching, Larsen-Freeman, 2006). After receiving relevant instruction and training, while the peer-assessment group was engaged in peer-assessment activities for four sessions, the self-assessment group was busy with self-assessment tasks. An achievement post-test (with phi dependability index of .90) was used to gauge performance differences at the end of the course. The application of ANCOVA indicated that the peer-assessment group outperformed the self-assessment group significantly, $F(1,37) = 7.13$, $P = .01$. Further findings and implications are discussed in the paper.

© 2012 The Authors. Published by Elsevier Ltd. Selection and peer-review under responsibility of ALSC 2012

Keywords: Alternative assessment, Peer-assessment, Self-assessment, Iranian university EFL students, Course achievement

1. Introduction

It is generally acknowledged that in any setting, where some systematic programs are practiced, there is a necessity for evaluation, in order to check their desirability and effectiveness. The educational systems

* Corresponding author. Tel.: +989198224704.
E-mail address: z.abolfazli1986@gmail.com; k.sadeghi@urmia.ac.ir
are not an exception. Traditionally, in a classroom setting, students’ learning was estimated through summative tests made by either the teacher him/herself in the specific school or institute or by the specialists (sometimes with a nation-wide scope), regardless of individual students’ needs and differences. However, as Jafarpour (1991) points out, if we are to increase the responsibility of the learner in EFL (English as a Foreign Language) study programs, testing procedures need to be adjusted.

A distinction is made by Brown (2004), between testing and assessing in that tests are formal procedures, usually administered within strict time limitations, to sample the performance of a test-taker in a specified domain. Assessment, however, is in general, an ongoing process, i.e. not limited to just one administration like the final achievement or summative tests at the end of the term, and it is wider in scope than formal tests in that it can measure learners’ learning along quite diverse ways; for instance, when a learner offers a comment or responds to a question, the teacher subconsciously makes an estimate of that learner’s state of knowledge.

Since the turn of the century, and in line with the developments in language teaching and learning, in most of which learners are taken as the authorities for their own learning, and as learner autonomy is gaining more importance, the responsibility of learning assessment is also shared among the learners themselves. As McNamara (2000) claims, more commonly achievement tests are more easily able to be innovative and to reflect progressive aspects of the curriculum and are associated with some of the most interesting new developments in language assessment in the movement known as alternative assessment.

Self-assessment is among one of the alternative methods of assessment which, as defined by Henner-Stanchina and Holec (1985), is the process whereby learners simultaneously create and undergo the evaluation procedure, judging their achievement in relation to themselves against their own personal criteria, in accordance with their own objectives and learning expectations. Sally (2005) states that self-assessment is not only important because it shows us what students know and do not know, but also because it provides students and teachers with feedback. Self-assessment is a key factor in authentic assessment and students’ empowerment because it provides students with the opportunity to reflect objectively on their own accomplishment and learning. Brown (2005) and Harris (1997) believe that self-assessment is an effective means of developing independent learning and helpful in changing perceptions of learning, i.e., in terms of skills and performance as opposed to amassing linguistic knowledge.

Another frequently adopted alternative assessment method in which students are actively involved in the appreciation and appraisal of learning is peer-assessment, as this is closely embedded in and aligned with students’ efforts during the instructional process (Shepard, 2000). According to Topping (1988), it is an arrangement in which individuals consider the amount, level, worth, and quality of success of the products or outcomes of learning of peers of similar status. In peer assessment, students learn from each other by means of receiving and giving feedback. Many students have reported that peer-assessment facilitates their learning (Ballantyne, et al., 2002).

2. Review of Literature

Barbera (2009) and Barrett (2007) state that authentic assessment involves student engagement in the evaluation process by using authentic evidences of learning processes and outcomes. Barbera (2009) and Sadler and Good, (2006) assert that self-assessment is particularly beneficial for students, since it not only keeps them involved, interested and highly motivated in the process, but also encourages self-reflection and responsibility. It is believed that when undertaking self-assessment, students are provided with opportunity to self-reflect on learning achievements, and to regulate their academic performance and
assessment behaviours (Barrett, 2010; Oskay et al., 2008). In scoring one’s own work, a learner will notice his/her strengths and weaknesses and have better understandings of the subject matter, which allows follow-up modification and improvement to be made (Chang, 2008).

One of the key arguments for self-assessment is that it provides “an effective means of developing critical self-awareness” (Nunan, 1988, p. 116). A resulting advantage of this is that learners are better able to set realistic goals and direct their own learning. A second argument is that in a learner-centred approach, by definition, learners need to be involved in “all the processes of learning, including the process of evaluation” (Little, 2005, p. 322). As it is impossible to teach learners everything they need to know and as learning does not stop outside the classroom, therefore it is of value to teach learners skills that they can transfer to other learning situations including skills in self-evaluation. While teachers may acknowledge the usefulness of learner-directed assessment, they continue to voice concerns on lack of appropriate self-assessment (Ekbatani & Pierson, 2000; Little, 2002). They remind us that one main concern is the learner’s capacity of assessing his or her own proficiency with any accuracy. However, it has been suggested that training helps and is in fact necessary, since self-assessment depends on a complex set of skills. Evidence also clarifies that accuracy is increased when self-assessment is in relation to clear descriptors and related to specific experience (Council of Europe, 2001). Another concern relates to its implementation. The need for self-assessment to be practical in terms of time and resources is stressed by Harris (1997), who suggests that self-assessment become an integral part of everyday classroom activities, i.e. a part of the regular procedures for dealing with reading, writing, listening, or speaking.

Davies (2000) believes peer-assessment to be beneficial to the learning process. Sengupta (1998) believes that unless the perceptions regarding teachers’ roles are addressed, it is probable that little value will be attached to peer-evaluation, and consequently collaborative and autonomous learning may not become a reality. Peer feedback to some extent is more influential and powerful than teacher feedback; therefore, the integration of self- or peer-assessment will effectively increase students’ involvement, motivation and incentives (Barbera, 2009; Jenkins, 2004). Chen (2010), and Chang et al. (2012) state that peer-scoring enables students to observe other people’s work, which brings new ideas and encourages self-reflection and improvement. Therefore, more than self-assessment serves, peer-assessment benefits students by offering opportunity to observe and compare peers’ works, emulate the strengths and avoid the weaknesses of other people. It also helps students keep track of their fellow students’ learning outcomes, providing powerful impetus to make progress and perform better (Bouzidi & Jalllet, 2009; Yang & Tsai, 2010). Additionally, students become aware of the quality of their own work.

Research studies examining peer-assessment have revealed that it can work towards developing students’ higher order reasoning and higher level cognitive thought (Birdsong & Sharplin, 1986), helping nurture student-centred learning among undergraduate learners (Oldfing & MacAlpine, 1995), encouraging active and flexible learning (Entwhistle, 1993) and facilitating a deep approach to learning rather than a surface approach (Entwhistle, 1987; Gibbs, 1992). Brown (2001) found that through peer-assessment, students enhance their sense of competence and self-worth. Peer-feedback can act as a socializing force and enhance relevant skills and interpersonal relationships between learner groups (Earl, 1986). Cheng and Warren (1997) and Kwan and Leung (1996) conducted studies on peer-assessment and found that students generally mark their peers within a narrower range than the class teacher to the benefit of the weaker students and to the detriment of the more able ones, an observation which is ascribed to the reluctance on the part of students to mark their peers up or down. Cheng and Warran (1997) also reached the conclusion that both the teachers and the students in their study found the peer-assessment exercise beneficial in terms of developing students’ higher level cognitive thinking and facilitating a deep approach to language learning. Some studies have observed a positive connection between prior training
and accurate peer-assessment (Williams, 1992; Forde, 1996). Moreover, since the student perception of the instruction is important, it is reassuring that student’ perceptions about the use of peer-assessment are very positive (Struyven, et al., 2003).

To clarify what the story is in an EFL university context and to contribute to the growing body of work in the field, the present research was aimed to answer the following question:

*Is there any statistically significant difference between self- and peer-assessment on Iranian university EFL students' course achievement?*

The question was tentatively answered in the form of a null hypothesis as follows:

*There is no significant difference between self- and peer-assessment on Iranian university EFL students’ course achievement.*

3. **Method**

This study was conducted to find out the effect of self- and peer-assessment on Iranian university EFL students’ course achievement. It was carried out with undergraduate students of English Language and Literature at Urmia University, West Azarbaijan, Iran. To this end, two experimental groups of self- and peer-assessment were required.

2.1. **Design**

Since it was impossible to assign the students into groups randomly, the two experimental self-and peer-assessment groups were both intact classes. This study can therefore be regarded as following a semi-experimental intact comparison group design. However, the classes were randomly assigned into either the self- or peer-assessment group.

2.2. **Participants**

The students were 40 EFL Bachelors of Art degree (BA) students at Urmia University, within the age range of 20 to 22. There were 19 (13 female and 6 males) in the self-assessment group and the peer-assessment group consisted of 21 (17 female and 4 male) students. The data coming from those participants who missed none of the treatment sessions (including the pretest or the posttest) were included in the final analysis.

2.3. **Instruments**

The particular course of interest was *Teaching Methodology* course for which *Techniques and Principles in Language Teaching* (Larsen-Freeman, 2006) was used and it was a four-unit credit bearing course. At the beginning of the semester, a knowledge test (pretest) was applied to check the homogeneity of students based on their existing Teaching Methods knowledge. Then the groups undertook four series of assessments, every two or three weeks. Finally, a course achievement test was administered as the posttest. Both the pretest and the posttest were adapted from the archive of final term exams for this course available at Payame-Noor University and were parallel forms of each other. Being content valid to a satisfactory degree, they had a phi (lambda) dependability index of .9038, using the following formula:

<table>
<thead>
<tr>
<th>Table 1: Descriptive statistics for the posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
</tr>
<tr>
<td>40</td>
</tr>
</tbody>
</table>
Valid N (listwise) 40

(1): \( \phi (\lambda) \) dependability formulae

\[
\phi(\lambda) = 1 - \frac{1}{K-1} \left[ \frac{M_p(1-M_p) - S_p^2}{(M_p - \lambda)^2 + S_p^2} \right]
\]

\( \lambda = \) the cut-off expressed as a proportion \( (.75) \)

\( k = \) number of items \( (35) \)

\( M_p = \) mean of proportion scores (or the mean divided by the number of items) \( (\frac{16.77}{35} = .4793) \)

\( S_p = \) standard deviation of proportion scores \( (\frac{1.7}{35} = .0486) \)

\[ \phi(\lambda) = .9038 \]

2.4. Procedure

At the beginning of the term, all the groups had a pretest which measured the students’ existing knowledge of the specific course book. Then, in the self-, and peer-assessment groups, the students were trained on how to assess themselves as well as their peers, respectively. For example the students were informed that they will have an assessment every two units covered from the course book, and are free to construct items of any type. While in the self-assessment group, they were instructed to make, to answer and then to mark their papers themselves, in the peer-assessment group, the candidates were instructed to answer tests made by anonymous peers.

The students in the self-assessment group were required to make and bring to the class some questions based on every two units covered. They were reviewed in order to check any obviously faulty items. This procedure was repeated for the next two units. In the third session, although the students were expected to deliver their third two units’ assessments, they were also required to undertake the first assessment i.e., the assessment based on the first two units. Finally, each one of the students marked their own papers and their scores were recorded.

In the peer-assessment group, students were required to design tests at home based on every two units covered. When brought to the class, their papers were gathered, and students’ names were taken apart. After assigning each student with a code, the papers were distributed among their peers, taking care not to give any paper to its own designer. When students had finished answering, the papers were given to their designers for marking. Then students were allowed to review their papers to eliminate any mismarking. Having received an assessment every two units, the two experimental groups took four assessments during the term. Finally, at the end of the term, both groups took the posttest.

2.5 Data analysis

At the pretest stage, an independent samples \( t \)-test was applied which showed a significant difference in the mean score of the two groups. The results for each one of the four series of assessments were
compared again via independent samples $t$-tests. At the end, the results of the posttest were compared via Analysis of Co-Variance (ANCOVA).

### 4. Results

In order to check the homogeneity of groups on the pretest (the knowledge test), an independent samples $t$-test was employed whose results are reported in table 2. As the table shows the difference between the mean scores of the two groups was statistically significant at .01 level of significance.

#### Table 2: Inferential statistics for pretest scores

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-assessment</td>
<td>19</td>
<td>6.8316</td>
<td>2.14063</td>
<td>.49109</td>
</tr>
<tr>
<td>Peer-assessment</td>
<td>21</td>
<td>5.3924</td>
<td>1.37073</td>
<td>.29912</td>
</tr>
</tbody>
</table>

Levene’s Test of Equality of Variances

<table>
<thead>
<tr>
<th>F</th>
<th>Sig.</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
<th>Mean Difference</th>
<th>Std. Error of Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.029</td>
<td>.143920</td>
<td>.56279</td>
<td>.16470 – 2.57852</td>
</tr>
</tbody>
</table>

The following four tables (from table 3 to 6) indicate the performance of self- and peer-assessment groups in the four associated series of assessments. Table 3 shows the results of the analysis derived from the first assessment in both groups which is significant at .01.

#### Table 3: Inferential statistics for the first assessment

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>First assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer-assessment</td>
<td>22</td>
<td>8.2955</td>
<td>1.34438</td>
<td>.28662</td>
</tr>
</tbody>
</table>

Levene’s Test of Equality of Variances

<table>
<thead>
<tr>
<th>F</th>
<th>Sig.</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
<th>Mean Difference</th>
<th>Std. Error of Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.192</td>
<td>.90455</td>
<td>.36607</td>
<td>.16470 – 1.64439</td>
</tr>
</tbody>
</table>

The following four tables (from table 3 to 6) indicate the performance of self- and peer-assessment groups in the four associated series of assessments. Table 3 shows the results of the analysis derived from the first assessment in both groups which is significant at .01.

#### Table 3: Inferential statistics for the first assessment

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>First assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variances</td>
<td>1.761</td>
<td>.192</td>
<td>2.471</td>
<td>.018</td>
</tr>
</tbody>
</table>
Table 4 shows that the difference in the mean scores of the two groups on the second assessment was also significant. What table 5 renders is that the two groups' mean scores again differed significantly on their third assessment. In contrast to the previous three assessments which indicated that the mean scores of the self- and peer-assessment groups differed significantly, as the digits in table 6 imply, the difference in their mean scores was not significant statistically (.81) for the fourth assessment.

### Table 4: Inferential statistics for the second assessment

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>self-assessment</td>
<td>17</td>
<td>9.7353</td>
<td>.56230</td>
<td>.13638</td>
</tr>
<tr>
<td>Peer-assessment</td>
<td>23</td>
<td>8.0000</td>
<td>.20295</td>
<td>.42320</td>
</tr>
</tbody>
</table>

Levene's Test of Equality of Variances

- F: 25.535, Sig: .000

- t-test for Equality of Means

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.73529</td>
<td>.50754</td>
<td>Lower: .70784, Upper: 2.76275</td>
</tr>
</tbody>
</table>

### Table 5: Inferential statistics for the third assessment

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>self-assessment</td>
<td>19</td>
<td>9.7368</td>
<td>.56195</td>
<td>.12892</td>
</tr>
<tr>
<td>Peer-assessment</td>
<td>18</td>
<td>8.6528</td>
<td>1.73870</td>
<td>.40982</td>
</tr>
</tbody>
</table>

Levene's Test of Equality of Variances

- F: 17.545, Sig: .000

- t-test for Equality of Means

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.08406</td>
<td>.42003</td>
<td>Lower: .23135, Upper: 1.93678</td>
</tr>
</tbody>
</table>

Equal Variances assumed

- F: 17.545, Sig: .000

- t-test for Equality of Means

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.08406</td>
<td>.42962</td>
<td>Lower: .18887, Upper: 1.97926</td>
</tr>
</tbody>
</table>
Since the two groups differed significantly on their pretest scores, Analysis of Covariate (ANCOVA) was used to compare their posttest mean scores. The value of sig. in Levene’s Test of Equality of Error Variance was .36 which means the assumption of equality of variances was not violated. The ANCOVA results are presented in table 7 and since the sig. value is .01, the two groups differed significantly in terms of their scores on the posttest. The corresponding effect size is .16, which according to Cohen (1988), reveals a small effect. It is obvious from the same table that the sig. value of the covariate is .03 and it is significant statistically, so controlling for the independent variable i.e., type of assessment, there is a significant relationship between the covariate, the difference in the groups’ pretest mean scores, and the dependent variable, the posttest mean scores. However, partial eta squared .12 reveals that the covariate had a small influence on the posttest.

Table 7: Univariate Analysis of Variance (ANCOVA) for posttest, with group as the fixed factor

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>21.962a</td>
<td>2</td>
<td>10.981</td>
<td>4.465</td>
<td>.018</td>
<td>.194</td>
</tr>
<tr>
<td>Intercept</td>
<td>649.261</td>
<td>1</td>
<td>649.261</td>
<td>264.020</td>
<td>.000</td>
<td>.877</td>
</tr>
<tr>
<td>Pre test scores</td>
<td>12.538</td>
<td>1</td>
<td>12.538</td>
<td>5.117</td>
<td>.024</td>
<td>.121</td>
</tr>
<tr>
<td>group</td>
<td>17.539</td>
<td>1</td>
<td>17.539</td>
<td>7.132</td>
<td>.011</td>
<td>.162</td>
</tr>
<tr>
<td>Error</td>
<td>90.988</td>
<td>37</td>
<td>2.459</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11372.330</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>112.950</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since the two groups differed significantly on their pretest scores, Analysis of Covariate (ANCOVA) was used to compare their posttest mean scores. The value of sig. in Levene’s Test of Equality of Error Variance was .36 which means the assumption of equality of variances was not violated. The ANCOVA results are presented in table 7 and since the sig. value is .01, the two groups differed significantly in terms of their scores on the posttest. The corresponding effect size is .16, which according to Cohen (1988), reveals a small effect. It is obvious from the same table that the sig. value of the covariate is .03 and it is significant statistically, so controlling for the independent variable i.e., type of assessment, there is a significant relationship between the covariate, the difference in the groups’ pretest mean scores, and the dependent variable, the posttest mean scores. However, partial eta squared .12 reveals that the covariate had a small influence on the posttest.

Figure 1 is a visual representation of the two groups’ performance at pretest, four series of assessments and posttest. As the table below the figure shows, the groups differed significantly in their pretest scores as well as in the first, second and third assessments, and these differences mainly show that self-
assessment group outperformed peer-assessment group. It also outperformed peer-assessment group in the fourth assessment, but the difference was not statistically significant. Altogether, as the last column in the mentioned table shows, it was the peer-assessment group rather than the self-assessment group that performed significantly better in the posttest.

![Figure 1: results from independent samples T-test for pretest, four assessment sessions in self-, and peer-assessment, and posttest scores](image)

5. **Discussion**

Since the students in the peer-assessment group had a statistically significant over-performance on the posttest than did the students in the self-assessment group, it can be concluded that the peer-assessment method significantly improves students’ learning in comparison with self-assessment. This finding is mostly in line with the results of Chan et al.’s (2012) study on portfolio assessment in which they found that peer-assessment group had the highest mean scores (3.86) followed by self-assessment (3.82) and the teacher-assessment with lowest scores (3.74). As the results of independent samples t-test revealed, the low mean scores of peer-assessment group in the series of assessments may be due to their willingness to score their peers as authentically as possible. Similarly, Chang et al. (2012) and Sadler and Good (2006) reported that peer-raters are stricter than self-raters. Peer-evaluation of writing was also found to have a significant impact on the improvement of the student writers (Brown, 2001; Patri, 2002). The difference in the performance of students in the self- and peer-assessment group in the present research can be said to be comparable with Patri’s (2002) study on the influence of peer-feedback on self- and peer-assessment, where it was claimed that the behaviour of peer-assessment has been somewhat different from that of self-assessment. Consistent with the findings of the present research, Lin et al. (2001) found that students in the self-and peer-groups had different performances and stated that a possible reason for the difference is that self-assessment is based on a more lax scoring standard than peer-assessment. Sadler and Good’s (2006) study, supporting the findings of the present research, revealed that peer-based scores were lower than self-based scores, which may mean that peer-raters tend to under-grade while self-raters tend to over-grade.

However, the finding of this study that self-assessors used a more lax scoring standard than peer-assessors did is in sharp contrast to what Chang et al. (2012), Lin et al. (2001), and Sadler and Good (2006) concluded. In their study they found that the teacher-scoring was the strictest, and peer-scoring was the most lax, with self-scoring in between, showing that peer-raters tended to adopt more lax scoring
standards than self-raters did. The analyses of the four series of assessments in the present study revealed a lower mean score in the peer-assessment group than in the self-assessment group, so peer-assessors might have been more strict in estimating their peers’ real achievement and therefore somehow underestimating their true knowledge. On the other hand, Pond et al. (1995) and Falchikov (1995) found peers less strict in assessing each other. They defined this over-marking by peers as ‘friendship marking’ or ‘decibel marking’, and claimed that this could be because peers found it difficult to criticize their friends. However, the divergent outcomes above might probably have been due to various educational levels of students, assessment rubrics, different assessment procedures employed, the assessment environments, assessor trainings, no autonomous learning experience, etc.

What is suggested here for further research is to design studies to control for the validity and especially the reliability of the assessments. Sulzen et al. (2008) identified high levels of validity for alternatives in assessment, but low levels of reliability and concluded that increasing the number of raters was effective in reliability improvement. It is assumed that students would be more likely to sharpen their rating abilities when provided with sufficient practice which will in turn make a higher validity possible. It must also be considered that peer-scoring results, however, are likely to be negatively affected because the practice of peer-assessment requires considerable amount of time and effort (Ballantyne et al., 2002; Davies, 2000; Tsai & Liang, 2009; Yang & Tsai, 2010). Therefore, in self-and peer-assessment, teachers should carefully deal with the issue regarding scoring practices and scoring burden in order to avoid potential factors that might adversely influence assessment results.

Orsmond et al. (1997), Sullivan and Hall, (1997), and Woolhouse, (1999) cautioned that subjectivity may apply both to self- and peer-assessment practices. More guidance on the marking criteria should be given to ensure that all markers can apply previously agreed criteria in a consistent fashion. Clear marking criteria give students the opportunity to see how their marks have been calculated. Adams and King (1995), and Pond et al. (1995) concluded that to enable students to perform peer- and self-assessment effectively, they need training and experience. Students must be given adequate training and practice in peer-assessment in order to minimize potential inconsistencies associated with subjectivity (Freeman, 1995). Patri (2002) suggests that when assessment criteria are firmly set, peer-evaluation enables students to judge the performance of their peers in a manner comparable to those of the teachers.

This study can be replicated a) at other levels e.g., MA, b) with other fields of study e.g., Engineering, c) by designing criteria for estimating item difficulties for the two groups, d) with other courses, e) with differing numbers of interim assessments, f) adding gender as moderator variable, g) with a focus on the consistency among the scores with the final achievement test, h) considering age and prior experience, i.e., low vs. proficient students assessing members of opposing proficiency level, i) comparing the items made by the students in each group from one session to the next, j) considering the possible effect of teacher differences by e.g., interviewing them before and after the study, and k) anticipating ways especially in the self-assessment group to encourage serious study.

Patri (2002) believes that since self-assessment and peer-assessment have been viewed as having significant pedagogic value, through their application, teachers could then focus more on enhancing their teaching techniques. The results may be used in all educational centres, with direct and indirect implications and applications for teaching, learning, material development, syllabus design, and test development.
6. Conclusion

This study was an investigation into the effect of self- and peer-assessment on Iranian university EFL students’ course achievement. The results of the analyses revealed that there was a significant difference in the course achievement of students in the self- and peer-assessment groups, with the latter outperforming the former. This over performance of students in the peer-assessment group on the posttest was contrary to what was expected from the analyses of the four series of assessments in all of which it was the self-assessment group which over performed the peer-assessment group. This observation implies that the students in the self-assessment group might have tended to be lax in assessing themselves as well as in scoring. Students in peer-assessment group were probably more willing to strictly assess their peers, which might have led to their underestimating their peers. So although they had low means in the four series of assessments, they performed much better on the posttest. Generally, for decision making on as accurate information as possible, it is suggested that none of self- or peer-assessment scores should account for a large proportion of students’ semester grades. However, since peer-assessment practices may lead to much higher levels of course achievement than self-assessment procedures (although this needs to be researched further), it is recommended that more weight be placed on the former in an alternative assessment programme.

7. References


