Advances in Language and Literary Studies
Vol. 3 No. 1; January 2012

# Self-assessment and Peer-assessment in an EFL Context

Morteza Yamini
(Islamic Azad University, Shiraz Branch, Iran)
Soheila Tahmasebi
(Islamic Azad University, Abadan Branch, Iran)

doi:10.7575/aiac.alls.v.3n.1p.49

## Abstract

Salient in an EFL teaching context is students' dissatisfaction with their final scores especially in oral courses. This study tried to bridge the gap between students' and teachers' rating system through alternatives to existing measurement methods. Task-based language assessment has stimulated language teachers to question the way through which students' language knowledge is assessed. Three groups of university students majoring in translation participated in this study. Two groups received rater instruction, but the control group did not. The assessed tasks were students' oral productions in Reading Comprehension I. Each oral production was assessed three times: by the speakers, by the peers, and by the teacher. The correlation of self-peer assessments and teacher assessments were estimated. Their performance on oral production of Reading Comprehension II was also analyzed and discussed and eventually compared with that of the control group to check the effects of rater instruction on learning.

## Introduction

New trends in the paradigm of teaching with their emphasis on the inevitable role of learners in the center of learning processes have positively affected the evaluation methods and eventually proposed alternatives in assessment (Brown & Hudson, 1998; Brown, 2004). As the name implies, alternatives in assessment bring about some changes in the process of evaluation and assessment, and involving students in the decisions made about them is among these changes. Some important characteristics of the new paradigm, specified by Brown and Hudson (1998), are as follows:

1. It requires students to perform, create, produce or do something, which is meaningful to them.

2. Students use language in real contexts and are constantly involved in assessing what they normally do in their classes, not in one session or at the end of a term.
3. Assessments are not intrusive and are welcome by the students since they get along in everyday activities.

Self-assessment and peer assessment as two alternatives of assessment have been dealt with by different authors with different perspectives. Brown and Hudson (1998) regard autonomy and intrinsic motivation as two theoretical principles underpinning self- and peer-assessment and consider cooperative learning as an extra asset of peer assessment. However, they also mentioned some drawbacks. Subjectivity is an important matter, which causes students to be too harsh and underestimate their abilities or be too lenient and overestimate themselves (Jafarpur & Yamini, 1995). However, Bailey (1998) concluded that there were correlations between self-rated oral production and scores on the Oral Proficiency Interview (OPI).

Not only through self-assessment can we involve students in decision making processes about their abilities, we can also use it as a helpful asset to enhance learning abilities. Gardner (1996) proposed a suggestion, which seems to go well with the EFL context in Iran. In assessing self-performance, he suggested using bilingual movies and news once with subtitles and once without. In this way, students not only assess their abilities but also provide feedback on their performance to improve their learning without receiving any pressure. Ellis (2003, p. 302) refers to “practical and educational advantages” of self-assessment and maintains that it is less time-consuming and less expensive to carry out. Moreover, he adds, where the purpose of instruction is to develop control over one's own learning, self-assessment is a useful tool for setting goal and providing reflective thinking.

Bachman (1990) mentions two studies and concludes that the way self-rating questions are framed affects the test takers' responses. Questions, which target linguistic abilities, do not represent students' language proficiency very well in contrast with questions related to students' actual needs and situations, which are better indicators.

Cheng and Warren (2005) investigated the reliability and potential benefits of incorporating peer-assessment into English language programs. The findings suggested that students had a less positive attitude towards assessing their peers’ language proficiency, but they did not score their peers’ language proficiency very differently from the other assessment criteria.

Students and teachers rated respective behaviors differently and interpreted oral and written language proficiency differently. These issues might tempt one to conclude that peer- and self-assessment results are not reliable; however, the prosperous effects presented through these innovations have other implications which diminish the dark side of the idea. The positive outcomes of self- and peer-assessment obtained from this experiment are outstanding. In fact, question number three of this study seeks for the advantages of integrating self- and peer-assessment in language programs, which is answered and discussed in the discussion section.

Some researchers have launched studies which consider the results of instructions on self-assessment as well as peer-assessment. Saito (2008) studied the effects of training on students who did peer-assessments in two experiments. The two experiments were almost the same except that in the second study the training hours increased. The result of the first experiment showed no significant difference between the treatment and the control group. The results of the second experiment revealed no difference, either, but regarding the quality and quantity of the comments provided, the experimental group scored higher. It could be concluded that instructing the participants to assess their peers had some positive effects on their overall production, if not directly affecting their assessing behaviors. Moreover, the results revealed that for the students to assess their peers, some instruction is needed since students' expectations and values could be different from those of the raters.

White (2009) launched a peer assessment (PA) study to determine student feelings about a student-centered assessment procedure, and to see if it was useful in promoting effective learning. The researcher allotted 30% of students’ final course grades to peer assessment scores of oral presentations. Students' perspectives on using peer assessment were positive, and the process promoted students' learning. The analysis also determined that student views are often congruent with views in the PA literature, despite the particular context of the investigation.

All the mentioned studies, though highly beneficial, did not target the effects of self- and peer-training on learning especially in a longitudinal study. Moreover, the results of the mentioned studies (Saito, 2008) are almost blurry and not precisely mentioned. This study tried to consider the difference between the self, peer and teacher assessment after students received rater training. Eventually, the effects of these alternative forms of assessment on learning were discussed through two oral presentation tasks done by three groups of the participants. Therefore, based on the above discussion the following null-hypotheses were made:

1. There is no significant difference between self, peer and teacher assessment of oral production after training.
2. Training on assessment has no significant effect on students' oral production.
3. Students have no positive views toward self- and peer-assessment.

## Method

### *Participants*

Sixty four EFL students majoring in English translation at Abadan Azad University participated in this study. During this experiment, they were in their first and second terms, taking Reading Comprehension I and II, respectively, as required by the educational office. They were considered as beginners. According to the university curriculum, they had to attend two-hour classes twice a week for thirteen weeks. They were pretested in their first terms of study to make sure that they were homogeneous. The students were divided into two groups. The control group consisted of 30 students and the experimental group consisted of 34 students. The source book and the teacher were the same for both groups.

### *Procedures*

At the end of the first term of their studies, students received four short texts and were asked to choose one of them and present it orally as part of their course requirement. In order to maintain students' interest, the selected texts had different topics including cell phones, love at first sight, great places to visit, and a ghost pilot. The students were asked not to memorize the sentences, but present the texts with the main ideas in mind. They were allowed to use their own sentences. This oral presentation, which counted for 20 percent of the final score, was filmed.

In the second term of the study, when students were taking Reading Comprehension II, they were divided into two groups randomly. The experimental group received a peer and a self rating scale and some instruction on how to score oral presentations for the first four sessions. After the fourth session, they were asked to assess one of their classmates who voluntarily presented the lesson of that day in order to practice assessing. The teacher and the presenter also did the same. Since they were beginners, they had problems in understanding some items of the assessment sheet and the teacher tried to help them in this respect. After rating, the scores assigned to each item, and the final scores were compared and students, as raters, put forward their reasons for giving a specific value.

The comments provided by the raters were divided into different areas ranging from metalinguistic to metapragmatic to even general humorous statements. This indicates how friendly and innovative such contexts were for the students who used to receive a single score at the end of the term without any feedback. The case for oral exams is even more severe since a lot of factors are intervening. This issue is taken up in the discussion.

When the students assured the teacher that they knew how to assess themselves and their peers, they were divided into two groups of 17, self-assessors and peer assessors. Out of the oral presentations of the first term, 17 were selected. They were bluetoothed twice; once to the speaker and once to a member of the peer assessors. All these procedures were done randomly. The teacher also assessed the same productions. Each student was asked to do the following tasks: 1) to transcribe the production, 2) to rate the production according to the rating scale, and 3) to provide some written comments. The students were allowed two weeks to hand in the papers. During these two weeks, extra rating practices were done for volunteers who wanted to be assessed by the teacher and students.

After two weeks, all the data were gathered, although some students brought their papers earlier. The results and the comments of each group were recorded for further analysis. The results of such findings have so far dealt with the first question of the study which targeted the difference between the peer, self and teacher assessments.

Since the second question of the study aimed at checking the effects of rater training on learning, the oral productions of Reading II which were similar to Reading I regarding the procedures were analyzed and discussed. However, the texts selected for this course were different; they matched the students' level of proficiency and had a readability of 12. There were 6 texts, and the students could choose one of them for their final oral production. They were handed out to the students one week prior to the final exam. Again, their production was recorded. The second question also compared the final production of the experimental group with that of the control group.

The control group benefited from the same teacher, used the same textbook, and had equal class hours as the experimental group did. However, they were different from the experimental group in that they did not receive any rater training and did not practice the assessment tasks that the experimental group did. To have an equal number for both groups, 17 students were asked to study the 6 texts presented to the experimental group and were asked to choose one of them for their oral production. They participated voluntarily. Their production was recorded and assessed according to the rating scale which was used for the experimental group. The scores were recorded to be compared with those of the experimental group.

### *Instruments*

1. Four texts with a readability of 10 were selected to be used as input. Each student chose one of them for their final oral production. The level of the readability was decided upon after analyzing students' performance on pretest, their class performance in general, and according to the level of the textbook that they were studying.
2. Students' oral production for Reading I was used in this study to be assessed by the participants in the study.
3. A rating scale assessing content, organization and fluency, each of which was divided into four categories ranging from very poor to excellent was used. Three scores were given to the same oral production using this scale; the speakers, their peers and the teacher assessed the same product.
4. Six short texts with a readability of 12 were selected whose length was about 10 to 12 compound or complex sentences. They were selected from Arco TOEFL. The reason for selecting these texts was that students in both groups had practiced similar texts during their class hours. These texts were used for both the control and the experimental groups.
5. Students' comments regarding the experience were considered as useful information, which shed some light on advantages and drawbacks of this study.

### *Analysis*

To find out the relationship between the different types of assessment correlation coefficients were found. The results are presented in Table 1.

**Table 1: Correlation coefficients between the different assessments**

| | | peer | self | oral production |
|---|---|---|---|---|
| teacher | Pearson Correlation | .692(**) | .298 | .890(**) |
| | Sig. (2-tailed) | .002 | .245 | .000 |
| | N | 17 | 17 | 17 |
| peer | Pearson Correlation | | .422 | .549(*) |
| | Sig. (2-tailed) | | .092 | .023 |
| | N | | 17 | 17 |
| self | Pearson Correlation | | | .264 |
| | Sig. (2-tailed) | | | .306 |
| | N | | | 17 |

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

As Table 1 shows, self-assessment was not significantly correlated with any of the assessments. It seems that the students had not learned to assess themselves properly. When one considers the means, it becomes clear that the students had assessed themselves higher than their own peers, whereas their teacher's assessment was lower than self-assessment and higher than peer-assessment. This becomes clear when we consider the means in Table 2. The teacher's evaluation of the oral productions with the mean of 16.29 falls between the peer- and self-assessment whose means are 14.82 and 17.18, respectively.

**Table 2: Mean and SD of different assessments**

| | N | Mean | Std. Deviation |
|---|---|---|---|
| peer | 17 | 14.8235 | 5.07734 |
| teacher | 17 | 16.2941 | 4.41255 |
| self | 17 | 17.1765 | 5.55917 |
| oral production | 17 | 18.5882 | 4.56972 |
| Valid N (listwise) | 17 | | |

In order to see if the instruction on peer- and self-assessment had an effect on the oral production of the participants, the performances of the control and experimental groups were compared through an independent-samples t-test. The results are displayed in Tables 3 and 4.

**Table 3: Group statistics**

| | group | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| oral production | experimental | 17 | 18.5882 | 4.56972 | 1.10832 |
| | control | 17 | 16.2353 | 3.25057 | .78838 |

**Table 4: Independent-samples t-test on oral production of control and experimental groups**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | |
|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. | Mean Difference |
| oral production | Equal variances assumed | 2.008 | .166 | 1.730 | 32 | .093 | 2.35294 |
| | Equal variances not assumed | | | 1.730 | 28.89 | .094 | 2.35294 |

As Table 4 shows, there was no statistically significant difference between the performance of the two groups in oral production.

## Discussion

Concerning the difference between peer, self, and teacher assessment after training, one can say that self-assessment showed no improvement, as it did not correlate with any of the variables in Table 1. However, peer-assessment showed a positive correlation with teacher assessment ($r = .692$, $p < .01$). This indicates that there was a 48% agreement between the teacher's assessment and students' peer-assessment. The null hypothesis can be rejected for peer assessment, but it should be retained for self-assessment. The students are unduly lenient towards themselves, but they do not show the same amount of leniency towards their peers. To remove or at least diminish this leniency, it seems that more hours of instruction are required to raise students' consciousness towards their performance. In other words, teachers should provide students with some feedback to help them assess themselves properly.

With regard to the effect of training on oral production, it has to be said that the mean difference between the two groups of control and experimental was not statistically significant ($t_{32} = 1.73$, $p =.093 > .05$). Therefore, the null hypothesis has to be retained. However, it should be noted that both groups have meaningfully improved in their oral production. This is clear from the assessment made by the teacher who showed a high correlation between her two evaluations ($r = .89$, $p< .01$). To further confirm the point, we ran a matched t-test between the two evaluations and the results are presented in Table 5. As the table shows, the mean difference is significant at the .01 level ($t_{16} = 4.474$, sig. = .000).

**Table 5: Matched t-test on teacher evaluation of 1st and 2nd oral productions**

| | Paired Differences | | | t | Df | Sig. |
|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Std. Error Mean | | | |
| teacher - oral production | -2.29412 | 2.11438 | .51281 | -4.474 | 16 | .000 |

The answers to questions 1-2 were provided statistically while the third question of the research had to be answered by means of qualitative data gathered through an interview and a questionnaire. The students' and teacher's experience in the control and the experimental groups was spectacular and helpful. Some of the important comments, which endorse the self and peer-assessments as alternatives in assessment, are presented below with the drawbacks following afterwards.

1. The atmosphere of the classroom changed. Students were no longer purely receptive in a reading course as they would in the traditional and mainstream methods. They were active and performed some tasks which were meaningful for them. In other words, they used language to do some real world tasks related to their studies.
2. The scores that they received and showed their success or failure were no longer vague to them. They knew why they scored low in oral production of Reading I. They confessed that they would no longer question the scores that their peers and they got.
3. After training and assessing themselves and each other, students were more cautious in delivering speech. They tried to consider the categories, which were mentioned in the rating scale; they avoided repeating the same ideas, and tried to speak clearly. That is, they monitored their production with some points in their mind.
4. The affective climate was incomparable to other classes experienced by the teacher. The students were friendly to each other and felt close to the teacher; they were doing what the teacher had been doing after all. They played verbal jokes and changed the frozen climate, which is typical for EFL classes into an attractive and funny climate. One of the students describing her own production said, "If you could understand what I said for Oral Production I, you are smart; I couldn't get a word of what I produced!" Another student describing his friend's performance said he would use it to go to sleep easily; it was a kind of lullaby!
5. A feeling of autonomy (during self-assessment) as well as cooperation while assessing their peers developed among students to evaluate their own abilities. As students, or after they graduate, they do not always have a rater within their reach to assess their

abilities and they should shoulder the burden of assessing themselves and their friends.

## Limitations and suggestions for further research

1. The number of the participants in this study was rather limited due to the fact that it was a classroom research and involving larger numbers could endanger the control of the teacher during the process. Involving greater number of participants in similar studies may hopefully end in results that are more valid.
2. Although this study continues for about nine months, extending this time and involving other factors such as sex and personality traits could yield in results that are more satisfactory.
3. Although the course of study to which this experiment was named Reading Comprehension just some part of the class hour and small percent of final score was devoted to oral production. Implying alternative assessments like self- as well as peer-assessment for other courses like speaking and even writing with this framework is also promising.
4. Not only self- and peer-assessments but other forms of assessments like portfolios, computer assisted language assessments are also useful hints to view language learning and evaluation through innovative methods.

## References

Bachman, L. F. (1990). *Fundamental Consideration in Language Testing.* Oxford: Oxford University Press.

Bailey, K. M. (1998). *Learning about Language Assessment: Dilemmas, Decision and Directions.* Cambridge, MA: Heinle & Heinle

Brown, H. D. (2004). *Language Assessment: Principles and Practices*. Longman Publishing.

Brown, J. D., & Hudson, T. D. (1998). *Criterion-referenced language testing and assessments: A teacher's guide.* Unpublished manuscript, University of Hawaii at Manoa.

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

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford: Oxford University Press.

Gardner, D. (1996). Self-assessment for Self-learners. *TESOLl Journal, 6*, 18-23.

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

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581.

White, E. (2009). Student perspectives of peer assessment for learning in a public speaking course. *Asian EFL Journal*, *33*(1), 1-36.