REDUCING SUBJECTIVITY IN SCORING A TEST
BY MEANS OF SELF-, PEER AND TEACHER ASSESSMENTS

Kusumarasdyati
Universitas Negeri Surabaya
kusumarasdyati@unesa.ac.id

received 13 January 2020; accepted 20 January 2020

Abstract
A common drawback of a subjective test lies in scoring, which may result in scores that are too high or too low. This problem occurred in one of the classes that the author taught in the English Department of Surabaya State University (UNESA) when two learners expressed their doubts about the objectivity of the author in giving scores. To overcome the problem, action research was conducted. The subjects of the study were twenty-four undergraduates taking the elective course Teaching English to Young Learners (TEYL). At the end of the semester, they took an oral exam to evaluate their understanding of the materials. During the exam, they had to assess themselves (self-assessment) and their classmates (peer-assessment). At the same time, the author as a lecturer also assessed them (teacher assessment). The results showed that the teacher assessment yielded the lowest mean (1.63), whereas the mean of scores obtained from the peer assessment was higher (1.76). It was the learners themselves who overestimated their knowledge about TEYL, resulting in the highest mean (1.79). The mark of each undergraduate was the average of the scores they got from the self-, peer and teacher assessments. In this way, the subjectivity of the teacher in assessing the undergraduates could be eliminated and fairness in determining the scores could be improved.

Keywords: scoring, subjectivity, self-assessment, peer assessment, teacher assessment

Introduction
Language tests could be classified into subjective and objective tests with respect to the scoring procedures (Bachman, 1995; Heaton, 1990; Mehren & Lehmann, 1991). An objective test contains a number of items, each of which has only one possible correct answer. Regardless of who scores these tests, the results remain the same. A subjective test, on the other hand, consists of fewer items. Each item requires a lengthy answer and the examiner subjectively determines whether this answer is correct or not, resulting in several degrees of correctness. The subjectivity of the examiner often affects the reliability of the test as the score may become lower or higher when the same answer is scored by another examiner (Scheeren et al, 2003). As a consequence, the learners may feel disadvantaged when they think they deserved better scores. For this reason, the examiner takes a number of actions to decrease the subjectivity in scoring the test, such as detailing
which aspects of constructs to be tested, devising a rubric that sets clear criteria of correct answers and having more than one rater score the test (Weir, 2005).

The problem of subjectivity in scoring an essay test also occurred in one of the classes that the author taught in the English Department of Universitas Negeri Surabaya (UNESA). At the beginning of the semester, two students expressed their concern that the author would be too subjective in giving grades to them. They attended another class that the author taught in the previous semester and thought they deserved a better grade than what they actually got. To overcome this problem, the author did an intervention by using many raters to score the test, hence decreasing the subjectivity in scoring. Rather than employing other lecturers as the raters, the author used self-assessment, peer assessment and teacher assessment in the exam at the end of the semester.

Self-assessment refers to the judgment made by a learner about his or her own performance in a test on the basis of certain criteria (Kleppin in Tassinari, 2015). By making such a judgment, the learner becomes more reflective and is aware of his or her own strengths and weakness, then he or she makes decisions about what and how to learn to improve themselves (Tassinari, 2015). Additionally, favorable results of self-assessment affect the learner’s confidence and increase his or her motivation to learn further (Gardner, 2000). In this way, the learner develops more autonomy in pursuing knowledge (Brown, 2004).

Peer assessment means “an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status” (Topping, 1998, p.250). In other words, it is an assessment about a learner done by fellow learners. Peer assessment is beneficial because it provides an opportunity to the assessing learners to be involved in the process of assessment, resulting in a good understanding of what assessment is and how it should be conducted (Benson, 2015). By critically examining other learners’ language performance, the peers could perform a similar task about their own performance more easily. Peer assessment, therefore, is a useful practice for learners to conduct self-assessment (Everhard, 2015).

It has been a common practice for teachers to use self-assessment, peer assessment and teacher assessment in evaluating the learners due to the above benefits. Empirical evidence on this issue has also been sought for the last few decades, and research showed that the good learners generally underestimated their performance in the self-assessment, whereas the poor learners overestimated theirs (Suzuki, 2015; Matsuno, 2009). Peers might assess the learners more leniently (Matsuno, 2009), but occasionally the latter scored themselves higher than former (De Grez et al, 2012). Despite the subjectivity of the learners, the peers and the teacher in scoring the test, it has been empirically proven that the highest improvement occurred when the self-assessment was performed simultaneously with peer and teacher assessments (Birjandi & Tamjid, 2012).

Taking this empirical evidence into account, the present paper aims to (1) describe the procedures of using self-, peer and teacher assessments concurrently to reduce subjectivity in scoring an oral test, and (2) find out whether there was a difference between the scores of self-, peer and teacher assessments. First, the methodology of the research will be explained in the next section. Afterward, the
following section will report the findings and elaborate on the discussion about these findings.

**Method**

Approached qualitatively and quantitatively, the present study used action research as the research design in order to overcome the problem in the form of subjectivity in scoring. The research subjects included forty-two undergraduates taking the course *Teaching English to Young Learners* (henceforth, TEYL) in the English Department of UNESA. They were selected purposively out of three TEYL classes because only the students in this class had a problem, i.e. casting doubt of the objectivity of the lecturer in scoring the exams and assigning grades.

To collect data from this class, two instruments were used, i.e. observation and a test. The observation was conducted to record the process employing the self-, peer and lecturer assessments. Since the lecturer participated as the examiner during the assessment, the observation was participatory in nature. In addition, the participatory observation, an oral test of TEYL was also administered. The test was a subjective one, in which each item consisted of a question that the students had to answer orally in length. The detailed account of the test administration will be presented in the next section.

Data collection took place in one day, i.e. when the mid-term exam was administered in October 2016. The qualitative data from the observation were analyzed by organizing, coding and interpreting them. The quantitative data in the form of test scores were analyzed by means of the mean to compute the average of scores and the ANOVA to find out the difference between the scores.

**Findings and Discussion**

The simultaneous use of self-, peer and teacher assessments turned out to be effective in solving the problem about subjectivity in scoring and giving grades in that particular TEYL class. In this section the procedures of conducting such assessment will be detailed, followed by the results of computation that demonstrate how the subjectivity could be reduced.

The procedures of the assessment began by giving clear instructions to the students so they knew what they could and could not do during the exam. Then the scoring sheet was administered to each student. It contained a table listing the name of the students and containing two columns where they wrote the scores for themselves and their peers (Figure 1). In this paper, pseudonyms replaced the real names of the students.
Scoring Sheet

Instructions: Assess your friends' and your own performance in the exam. Write the scores (between 0 and 3) in the appropriate columns.

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Item no. 1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. An excerpt of the scoring sheet

The lecturer also had a copy of the scoring sheet in her hand for the purpose of teacher assessment. She explained to the students that the score of each item was determined by subjectively judging the correctness of the answer. A rubric was shown to the students and served as the guidelines in giving scores (Table 1).

Table 1. Rubric for scoring

<table>
<thead>
<tr>
<th>Scores</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>The answer is absolutely correct.</td>
</tr>
<tr>
<td>2</td>
<td>The answer is almost correct.</td>
</tr>
<tr>
<td>1</td>
<td>The answer is a little bit correct.</td>
</tr>
<tr>
<td>0</td>
<td>The answer is absolutely incorrect.</td>
</tr>
</tbody>
</table>

The scores ranged from 0, which means the answer was absolutely incorrect, to 3, which means the answer was absolutely correct. If the students judged the answer contained inaccuracy to a low degree, they gave a score of 2. However, if the inaccuracy was greater, they gave a score of 1.

The seat arrangement was organized in such a way that the examiner and the examinees could see and hear each other clearly (Figure 2). This was essential because it was an oral exam. Failing to hear the questions or answers clearly would affect scoring, and decrease the reliability and the validity of the test.

Figure 2. Seat arrangement
The lecturer faced three test-takers, and the peers faced the lecturer and the test-takers. First, the lecturer asked a question orally to a test-taker, who answered the question also orally. Immediately after that, the lecturer told the answer key to the test-taker and the peers so that all students knew the best answer to the question. Knowing the correct answer, the test-taker did self-assessment by writing the score for himself or herself on the scoring sheet. At the same time, the peers did peer-assessment by writing the score for the answer that the test-taker gave. The lecturer also assigned a score on the scoring sheet, and this was the teacher assessment. Next, the lecturer asked another question to the test-taker, and the above procedures were repeated so that two scores were obtained for this particular test-taker.

After the first test-taker answered two questions, the second and the third test-takers also got two questions, resulting in two scores for each of them. The scoring procedures remained the same as the above. When scoring the third test-taker was completed, three students who sat in front of the lecturer moved to the left side of the classroom where the peers sat. They became the peers, and three of their classmates moved to the seats before the lecturer. They became the next batch of test-takers, and the aforementioned procedures were repeated until all students got their turn in answering the questions and assigning scores. The lecturer collected the scoring sheets from the students and it marked the end of the oral exam.

The self-, peer and teacher assessments described above resulted in three sets of scores. To find out whether there was a significant difference between them, ANOVA was applied and the results of the computation are presented in Table 2.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>N</th>
<th>$\bar{X}$</th>
<th>$\sigma^2$</th>
<th>F</th>
<th>$F_{0.05}$</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>24</td>
<td>1.79</td>
<td>0.45</td>
<td>3.13</td>
<td>Not significant</td>
<td></td>
</tr>
<tr>
<td>Peer</td>
<td>24</td>
<td>1.75</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>24</td>
<td>1.63</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The above numbers were shown to the students so that they have a thorough understanding of how the results of the assessment were and what they meant. The highest mean was obtained from the scores of self-assessment ($\bar{X}$=1.79), whereas the lowest was from the scores of teacher assessment ($\bar{X}$=1.63). The scores of peer assessment yielded a mean which was slightly lower ($\bar{X}$=1.75) than the mean of self-assessment scores. This finding was consistent with the result of the study conducted by De Grez et al (2012), who found that students tended to rate themselves higher than the teacher.

The variance of the self-assessment scores was also the highest ($\sigma^2$=0.45), indicating the widest spread of scores. Interestingly, there was a slight difference in the variance between the teacher-assessment scores ($\sigma^2$=0.42) and the self-assessment scores. The scores of peer assessment were more homogeneous than the other two sets of scores, as shown by the lower variance ($\sigma^2$=0.20).

To find out whether there was a significant difference between the means of self-, peer and teacher assessments, the statistical computation was performed by means of ANOVA. The result indicated that the difference between the means was not statistically significant ($F$=0.51). Thus, there was actually no difference between the scores of self-, peer and teacher assessments. A possible reason for the
insignificant difference was the use of rubric which guided the lecturer and the students in assigning scores. Simple as they might seem, the criteria set in the rubric helped the examiner and the examinees to make similar judgments about the examinees’ answers. Another reason was the answer keys provided by the lecturer immediately after the answers were articulated by the students. The answer keys served as a standard in determining to what extent a particular answer was correct, making the examiner and the examinees have very similar perceptions about the degree of correctness for each answer. In this way, the subjectivity in scoring could greatly be reduced, and—more importantly—the problem that prompted the author to conduct the present action research was solved.

Afterward, the mean of the scores that each student got from the self-, peer and teacher assessments was computed. The average of the three scores became his or her mid-term exam score. The results of the computation for the whole class are summarized in Table 3.

<table>
<thead>
<tr>
<th>Highest Score</th>
<th>Lowest Score</th>
<th>( \bar{X} )</th>
<th>( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.87</td>
<td>0.62</td>
<td>1.72</td>
<td>0.28</td>
</tr>
</tbody>
</table>

The average of the scores was 1.72. The highest score in the class was 2.87, while the lowest one was 0.62, resulting in a range of 2.25. Although the difference between the highest and the lowest scores was quite high, the computation of the variance obtained a value of 0.28. This means when all of the scores were taken into account, the variability of the scores was relatively low (cf. the variance of the self-, peer and teacher assessments in Table 2), suggesting that the scores were quite homogenous.

As mentioned above, all results of the computation and their interpretation were presented to the students after a few days. They responded to this transparency in assessment positively. They understood that—as the name suggests—a subjective test inevitably involved the subjectivity of the raters in scoring. They realized that they might have overestimated or underestimated their own ability when assessing themselves or others, and some procedures could be done to lessen the subjective judgment.

**Conclusion**

Subjectivity in scoring has been a persistent problem in scoring an essay test, and adult learners in the university were particularly concerned about it so action research was conducted to overcome this problem. In the mid-term exam, the assessment was conducted by not only the lecturer but also the students themselves and their peers, resulting in three scores for each student. The average of these scores was computed and became the mid-term exam score. These procedures reduced the subjectivity to a large extent because the assessment took into account the perspectives of three parties rather than relying on the judgment of the lecturer only. The students could understand and accept such a form of assessment, no longer casting doubt on its subjectivity. Thus, the problem of subjectivity was successfully overcome through action research.
References