# PEER ASSESSMENT IN UNIVERSITY LEVEL: A PRELIMINARY STUDY ON THE RELIABILITY

**Suwarni Wijaya Halim**
Universitas Bunda Mulia, Jakarta
Pos-el: suwarni@bundamulia.ac.id

## ABSTRACT

This research aims to find out the reliability of peer assessment as a form of alternative assessment and as a part of student-centered learning process. The data was taken from one of the 'Writing 3' classes in English Language and Culture Department, which consist of fifteen students. The students were taught how they should conduct peer assessment, what rules they should follow, and how they should utilize the provided scoring rubric before they assessed their peers' writing assignments. The writer, as the lecturer of the class, also conducted her own assessment using the same scoring rubric. The results of peer and teacher/lecturer assessment were compared and calculated using SPSS in order to find out whether the results of peer assessment have significant difference with the results of teacher/lecturer assessment. The result shows that there is no significant differences between the results of teacher assessment and peer assessment. This means that peer assessment could be implemented as a form of alternative assessment in writing classes.

**Keywords**: peer assessment, teacher assessment, reliability

## *ABSTRAK*

*Penelitian ini bertujuan untuk mengetahui reliabilitas dari penilaian sejawat (peer assessment) sebagai suatu bentuk dari penilaian alternatif dalam kegiatan mengajar dan sebagai bagian dari proses pembelajaran yang berpusat pada pelajar (student-centered learning). Data untuk penelitian ini diambil dari salah satu kelas 'Writing 3' dari program studi Bahasa dan Budaya Inggris yang terdiri atas 15 orang. Para mahasiswa diajarkan bagaimana mereka seharusnya melakukan penilaian sejawat, peraturan apa yang harus mereka patuhi, dan bagaimana mereka menggunakan rubrik penilaian yang telah disediakan sebelum mereka akhirnya menerapkan kegiatan penilaian sejawat dalam menilai tugas menulis teman-teman mereka. Penulis, yang juga merupakan dosen dari kelas tersebut, juga melakukan penilaian dengan menggunakan rubrik yang sama. Hasil penilaian dari penulis dan dari para mahasiswa dibandingkan dan dikalkulasi menggunakan SPSS untuk mengetahui apakah nilai yang dihasilkan dari penilaian sejawat dan penilaian guru memiliki perbedaan yang signifikan. Hasil yang diperoleh menunjukkan bahwa tidak ada perbedaan yang signifikan antara kedua bentuk penilaian. Hal ini menunjukkan bahwa penilaian sejawat dapat digunakan sebagai bentuk penilaian alternatif dalam kelas menulis.*

***Kata kunci**: penilaian sejawat, penilaian guru, reliabilitas*

## A. INTRODUCTION

Generally, when teachers or lecturers give assignments to the students, the process mostly follows the same cycle—the lecturers deliver assignments; the students do the assignments and submit them to the lecturers; the lecturers grade the assignments, give scores and feedback, and return the assignments back to the students—and the cycle goes on and on.

This type of learning can be perceived as a traditional, teacher-centered learning. The teaching and learning process is very much dependent on the role of teachers or lecturers. Teachers or lecturers are viewed as the sole authority and knowledge providers who are capable of transmitting knowledge to the students and as the judges who are able to deliver judgment on the quality of the students' works.

As the knowledge and theories of teaching and learning progresses, the traditional, teacher-centered learning receives more and more criticisms. One among many criticisms for this teaching style is the fact that the students are expected to be the passive party in the process of learning when they should have been encouraged to establish their autonomy, assert their independence, and actively pursue their educational goals. For the last fifty years, the experts have been trying to develop alternative ways of learning. Instead of just sitting and listening to the lecturers' explanation, the students are encouraged to learn independently—this method is referred to as student-centered learning.

Weimer (2013) explained that the distinctive features of teacher-centered and student-centered learning can be seen from the role of teachers and students. In teacher-centered learning, the teachers or lecturers act as the leader of the class and the expert in the field, and they provide explanation and transmit knowledge while the students just listen and take notes. In student-centered learning, the students are required to be active and independent individuals who take responsibilities for their own learning while the teachers or lecturers only act as the manager and the monitor in the classroom activities. In other words, the teachers or lecturers share their power and authority with the students so that the students can independently and autonomously manage their own learning (Dewi, 2015).

The distinctive features do not only stop in terms of teaching and learning process but also in terms of assessment. As mentioned above, in teacher-centered classroom, the assessment is usually conducted by the teachers or lecturers because they are perceived as the experts and have more than sufficient knowledge and experiences to deliver judgments when assessing students' assignments. However, in student-centered learning, the students are actively involved in the process of assessment.

In the writing classes in English Language and Culture Department at the writer's institution, the assessment on writing assignments has been largely dominated and conducted by the lecturers. The lecturers have the responsibilities to provide grades and give feedbacks and comments for the students' writing assignments. They need to read the writings composed by the students, scrutinize from minor to major details, and think of a way to give constructive and useful feedbacks for the students. This process obviously demands long time and much effort from the lecturers, and the process becomes even more daunting when the lecturers are assigned to teach writing classes with approximately twenty-five to thirty students.

The further problem with the teacher-centered assessment in writing classes is that the students do not seem to learn much from the grades and feedbacks provided by the

lecturers. One of the proposed solutions for this situation is by implementing alternative forms of assessment—in this case, student-centered assessment—in writing classrooms. The forms of student-centered assessment include self and peer assessment. This research would only focuses on the notion of peer assessment.

Topping (1998, p. 250) defined peer assessment as "an arrangement in which individuals consider the amount, level, value, worth, quality of success of the products or outcomes of learning of peers of similar status." In other words, in peer assessment, students assess and evaluate their friends' works and provide comments and feedbacks for them.

After reviewing some related literatures, the writer proposed an investigation on the implementation of peer assessment in 'Writing 3' class, specifically in terms of the reliability of the peer assessment because it is unknown whether the peer assessment would work well if implemented in 'Writing 3 class'.

The writer conducted this research because the writer would like to find out whether peer assessment could be a reliable alternative assessment to teacher assessment. This research attempted to answer the following research question: "How does the result of peer assessment compare with the teacher assessment?"

The objective of this study is to find out whether there are significant differences between the scores produced by the writer and the scores produced by the students. The result of the comparison would supply the writer with information to conclude whether peer assessment is reliable to be implemented in writing class.

In term of significance of the study, the writer hopes that this research would give benefits for the lecturers, particularly those who teach writing class, so they can acquire more information about peer assessment before they consider using peer assessment in assessing their students' writing assignments. Furthermore, the research is hoped to benefit the students so that they can improve their writing quality and their writing skills.

The study was conducted towards one out of two 'Writing 3' classes. The participants of this study are fifteen 4$^{th}$ semester students in that particular class. The writer chose this class because the writer was also the lecturer for the said class. It was, therefore, easier for the writer to explain the mechanisms of peer assessment, to provide sufficient scaffolding and opportunities for the students to learn about peer assessment and to ask questions, to provide the writing tasks, to monitor the implementation of peer assessment, and to discuss the results of peer assessment.

This research can be considered as a preliminary case study since the writer only observed and analyzed the implementation of peer assessment in a class of fifteen students, in which the writer was the lecturer. The students were carefully taught about the procedures and regulations in conducting peer assessment prior to the real implementation. The data to answer the research questions were gathered from all 15 students. This research utilized quantitative approach, and the writer used writing assignments as the source of data.

## B. THEORETICAL FRAMEWORK

### 1. Peer Assessment

Falchikov (1995, as cited in Sluijsmans, Dochy, & Moerkerke, 1998, p. 300) defined peer assessment as "the process whereby groups of individuals rate their peers." Similarly, Roberts (2006, p. 6) defined peer assessment as "the process of having the learners critically reflect upon, and perhaps suggest grades for, the learning of their peers." There are several points that need to be emphasized in this case. The first point is

that the actor doing the act of assessment is the student, and he or she is assessing the works of other students that have the same status as him or her. The second point is that the object of assessment is the result of learning—in this case, the peers' writing assignments. The third point is that the focus of peer assessment is to measure how the peers have successfully written a good piece of writing according to the agreed standard of measurement.

Over the years, there have been abundant studies regarding the implementation of peer assessment and its effects on the process of teaching and learning. Most previous research on peer assessment can be categorized into three broad issues, which are the issue of quality of peer assessment result, the advantages and disadvantages of implementing peer assessment, and the perception of the students about peer assessment.

In terms of the quality and trust-worthiness of peer assessment practice, Burke (1969, as cited in Freeman, 1995) stated that compared to self-assessment, peer assessment is more reliable. This conclusion was drawn after the research findings showed a very high rate of agreement among the peers and a high rate of agreement between results generated from peer assessment and teacher assessment. Freeman (1995) also conducted a study by comparing the results of peer assessment and teacher assessment in evaluating oral presentations and found no significant differences between the two assessments. Moreover, Topping (1998, p. 262) emphasizes that "peer assessment appears capable of yielding outcomes at least as good as teacher assessment and sometimes better." This implies that the results generated from peer assessment are highly trust-worthy and qualified. On a similar note, Ramon-Casas, Nuño, Pons, and Cunillera (2018) analyzed the inter-grader agreement and consistency of the scores, and in the end, they reported that peer assessment could provide highly valid and reliable result in assessing the writing tasks.

Further studies on the literature, however, reveal another results and interpretations. Cheng & Warren (2005), who conducted study on peer assessment in seminar, oral presentation, and written report, found that there was a significant difference between the results of peer assessment and teacher assessment in terms of evaluating oral proficiency and written proficiency. They investigated further and discovered that the perceptions that students had about oral and written proficiency were different from the perceptions that the teachers had. As a result of the difference in perception, the results generated from peer and teacher assessments were also significantly distinct.

In terms of benefits or advantages of implementing peer assessment, a research by Tsui & Ng (2000) shows that most students favored teacher's assessments and comments more than their peers' assessment and comments. Nevertheless, results of peer assessment brought positive effects. Because they were aware of the fact that their peers were going to read their writing, they would try harder to maintain the strengths and improve the shortcomings of their writing. Similar findings were also found in the research conducted by Min (2006) and White (2009). Min (2006) reported that after a careful training and scaffolding process on peer assessment, students were able to enhance their quality of writing and revising. Similarly, White (2009) found that peer assessment enhances the quality of learning in public speaking class. In addition, Rofiudin (2011) found that peer assessment can improve students' skills in writing descriptive texts and increase their interest on writing. Purnawan (2015) also asserted that the peer assessment does not only benefit the students but also the teachers. By conducting peer assessment, the teachers can minimize their subjectivity when assessing the students' writing assignments. Furthermore, Everhard (2015) mentioned that peer assessment can be a good practice and

starting point for self-assessment since peer assessment helps the students improve the skills and objectivity needed in the process of self-assessment.

Despite the fact that many benefits can be reaped from the implementation of peer assessment, several drawbacks are also discussed. Topping (2009) revealed that a number of teachers might be apprehensive and anxious in applying peer assessment in the classroom, especially in the case of summative assessment in which the final grades will be given. Topping (2009) also mentioned that a range of problems caused by "social processes" can also affect the result of peer assessment:

> Social processes can influence and contaminate the reliability and validity of peer assessments. Peer assessments can be partly determined by friendship bonds, enmity, or other power processes, the popularity of individuals, perception of criticism as socially uncomfortable, or even collusion to submit average scores, leading to lack of differentiation. (Topping, 2009, p. 24)

On a similar note, the research by Roberts (2006) revealed that students might also experience apprehension regarding the implementation of peer assessment since they are not confident that they have the necessary skills to evaluate and judge their peers' work and since they believe that assessment should be conducted by their teachers or lecturers.

In terms of perception on peer assessment, there has been a conflicting point of view. White (2009) examined the students' perspectives about the implementation of peer assessment in public speaking course by utilizing survey, and the results show that the students gave positive response to the implementation. Sumekto (2014) distributed questionnaires in the form of Likert Scale from 1-5 in order to find out the perception of university students in Muhammadiyah University of Purworejo. The findings show that 52 percent of the respondents responded favorably about peer assessment. It was also found that "the perception is supported by the lecturer's trust, assessment accuracy, and students' expectation" (Sumekto, 2014, p. 1137).

A completely different and interesting finding, however, was brought forward by Cheng & Warren (2005). They interviewed the students after the students conducted peer assessment for several times, and they found that the students were not confident and comfortable in assessing their peers' writing works. Some of the students also revealed that they did not feel adequate enough to judge and evaluate their friends' works since they believed they did not have enough capabilities to do so. Cho, Schunn, and Wilson (2006) investigated teachers' and students' perception about the reliability and validity of the grades and scores generated from peer assessment. They found that the teachers perceive the result of peer assessment as highly valid and reliable. On the other hand, the students think that the results of their own peer assessment as invalid and unreliable. The research conducted by Kaufman and Schunn (2011) also yielded similar findings. They conducted a comparative study on peer assessment by examining two situations. In situation one, the students conducted the peer assessment while the teacher also conducted teacher assessment on the same writing assignments. In situation two, the teacher did not conduct any assessment, so the assessment was fully conducted through peer assessment by the students. They found that the students had positive attitude on peer assessment when the teacher was also involved in the assessment (situation one). However, their perception significantly changed into negative when only the students conducted peer assessment (situation two). It was because the students did not think assessment conducted by their peers is fair, qualified, valid and reliable.

## 2. Reliability

Reliability is an important notion in the assessment. Harris (1969) included reliability as one of the criteria that an assessment must have, along with validity and practicality. Similarly, Brown (2004) also stated that reliability is a necessary component of an assessment, together with the notion of practicality, validity, authenticity, and washback.

Reliability can be defined as the degree of agreement in terms of generating the scores. An assessment can be considered as reliable if it generates similar and consistent results on different occasions by different people (Brown, 2004; Kubiszyn & Borich, 2013). For example, an assessor is tasked to assess one writing assignment and is instructed to assess the same assignment twice—in first week and second week. In order for the assessment to be considered as reliable, the scores produced in first-week assessment and second-week assessments have to be similar or at least, do not have significant differences. If both scores have proven to reach certain degree of agreement in the calculation process, the assessment can then be considered as reliable.

There are two most common types of reliability, i.e. intra-rater agreement and inter-rater agreement. Intra-rater agreement is similar to the situation described as an example in the previous paragraph. Only one assessor is involved in the process, and it aims to test mostly the credibility of the assessor. Inter-rater agreement, however, involves many assessors. Each of the assessors is tasked to assess certain amount of works, and the results from each assessor will be compared. If the results show agreement among the assessors, it can be concluded that the assessment is reliable.

## 3. Framework for Research

In this research, the writer focused on the notion of inter-rater agreement. It is because based on these previous studies, it can be concluded that the students are slightly anxious about their own and their peers' capabilities and fairness in conducting peer assessment. Taking the research conducted by Cho, Schunn, and Wilson (2006) and Kaufman and Schunn (2011) into account, the writer would conduct the teacher assessment together with the students who would conduct the peer assessment. The assessors, in this case, were the writer herself as the teacher or lecturer of the class and the students of writing class. Both scores from the writer and the students would be compared to determine whether or not peer assessment is reliable.

Moreover, taking the research conducted by Topping (2009) and Ramon-Casas, Nuño, Pons, and Cunillera (2018) into account, the writer would attempt to address the concern on subjectivity and problems that might be caused by social processes. One of the solutions put forward by the previous researchers is the use of a structured set of standards that the students can refer to when they are assessing their peers' works. The set of standards that the writer used in this research is the rubric by Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) (see Appendix) which have been used in numerous research for assessing writing tasks. Furthermore, before the implementation of the actual peer assessment process, the writer also made sure to provide sufficient scaffolding for the students by explaining and testing how the peer assessment would work and by giving the students ample opportunities to ask questions and discuss the technicalities of the process.

## C. METHOD

The writer employed quantitative approach in this study. This is due to the fact that research questions posed in this proposal need quantitative means to answer. The research question is about the reliability of the results of peer assessment conducted by the students in comparison to the results of assessment conducted by the writer as the teacher or lecturer of the class. This research question was answered by using SPSS ("IBM SPSS Statistics for Windows," n.d.) in order to find out whether the results of peer assessment by the writer and the students have significant differences or not. Based on the information gathered from the calculation, the conclusion on the reliability of peer assessment was drawn.

The data were gathered from 15 students in 'Writing 3' class. The writer chose this class because the writer is also the lecturer for the said class. It was, therefore, easier for the writer to explain the mechanisms of peer assessment to the students and to administer the research instruments in the class.

In order to collect the data, the writer executed several stages in order to ensure that the data collection was conducted successfully. The stages can be categorized into two major steps: Initial Preparation and Actual Implementation. In the Initial Preparation stage, the writer explained the procedures and guidelines of peer assessment before the actual implementation of the peer assessment. The writer showed how to conduct the assessment using scoring rubric by Jacobs et al. (1981) (see Appendix). The writer also explained the rules of giving constructive feedbacks to the students so that the students can give their peers helpful comments. Afterwards, the writer asked the students to try conducting peer assessment on their friends' writing assignments as a form of practice. The writer then monitored the class and provided help for the students when they need to.

After the short training course, the students were given writing tasks of various topics as outlined in the syllabus. The types of writing ranged from descriptive, cause-and-effect, and argumentative essays. Afterwards, the students were instructed to conduct the process of peer assessment. The assignments were randomly distributed to the students, and they had to assess the other students' writing homework based on the rubric. The writer then noted down the scores generated from peer assessment. This process of peer assessment was conducted again for several times with different writing assignments in order to generate sufficient number of data. The scores from peer assessment, along with the scores from the writer, who was the lecturer of the class, were later on used as the data to answer the first research question.

There are several steps in managing, organizing and analyzing the data. The data in the form of scores was calculated to find out the average value of the assessment by lecturer and the assessment by peers. The average scores were inputted into SPSS in order to find the results.

First, the writer would test the normality of the data in order to decide whether the distribution is normal or not. If the distribution is normal, the writer would use paired-samples t-test in order to identify whether there is significant difference between the writer and the students' results of assessment. If, however, the result of normality test shows abnormal distribution, the writer would use non-parametric procedures. In order to calculate the reliability, the writer would employ Wilcoxon signed-rank test, which is the non-parametric equivalent to paired-samples t-test (Corder & Foreman, 2009).

## D. FINDINGS AND DISCUSSION

As stated previously, while the students conducted the peer assessment, the writer as the lecturer of the class also conducted the assessment using the same scoring rubric by Jacobs et al. (1981) that the students used (see Appendix). The results generated from both peer and teacher/lecturer assessment would be compared later on. Listed in the table below are the scores that are generated from both peer assessment and teacher/lecturer assessment.

**Table 1. Results of Peer and Teacher/Lecturer Assessment**

| Name | Task 1 (Lecturer) | Task 1 (Peer) | Task 2 (Lecturer) | Task 2 (Peer) | Task 3 (Lecturer) | Task 3 (Peer) |
|---|---|---|---|---|---|---|
| Student 1 | 71 | 83 | 78 | 71 | 0 | 0 |
| Student 2 | 77 | 74 | 76 | 73 | 0 | 0 |
| Student 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 5 | 78 | 76 | 71 | 83 | 67 | 63 |
| Student 6 | 78 | 85 | 79 | 78 | 0 | 0 |
| Student 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 8 | 81 | 75 | 0 | 0 | 77 | 77 |
| Student 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| Student 10 | 87 | 88 | 91 | 89 | 93 | 91 |
| Student 11 | 80 | 80 | 77 | 78 | 84 | 89 |
| Student 12 | 73 | 70 | 70 | 66 | 65 | 75 |
| Student 13 | 0 | 0 | 78 | 77 | 0 | 0 |
| Student 14 | 78 | 78 | 75 | 76 | 79 | 83 |
| Student 15 | 90 | 98 | 90 | 86 | 94 | 98 |

**Table 2. Average Results of Peer and Teacher/Lecturer Assessments**

| Name | Average (Lecturer) | Average (Peer) |
|---|---|---|
| Student 1 | 50 | 51 |
| Student 2 | 51 | 49 |
| Student 3 | 0 | 0 |
| Student 4 | 0 | 0 |
| Student 5 | 72 | 74 |
| Student 6 | 52 | 54 |
| Student 7 | 0 | 0 |
| Student 8 | 53 | 51 |
| Student 9 | 0 | 0 |
| Student 10 | 90 | 89 |
| Student 11 | 80 | 82 |
| Student 12 | 69 | 70 |
| Student 13 | 26 | 26 |
| Student 14 | 77 | 79 |
| Student 15 | 91 | 94 |

As seen from Table 1 above, there were several students who got 0 (zero) for their assignment score. This was due to the fact that either the students did not submit their assignments or the students committed plagiarism. When the students failed to collect their assignments or when the students were caught plagiarizing, they would automatically get zero score for that assignment. This policy was introduced at the

beginning of the semester during the class introduction and re-emphasized during the assignment of each writing task. All the students, therefore, were aware of this policy and the consequences.

The data in Table 1 were then calculated for the average values. The result of the calculation was rounded so that there were no decimal values. The result can be seen in the table below.

A quick glance on Table 2 would show that the average results of both lecturer and peer assessments do not differ greatly. However, it is important to draw the conclusion based on proper calculation. Therefore, the average results were inputted into the SPSS in order to find out whether the data has normal or abnormal distribution.

The SPSS usually generates two types of statistics for the tests of normality, which are Kolmogorov-Smirnov and Shapiro-Wilk. According to Thode (2002), Kolmogorov-Smirnov is the most common statistical test used in testing the normality of the data. However, Shapiro-Wilk generates more powerful result and is considered more suitable for research with less than 50 data. The writer, therefore, would look at the result generated from Shapiro-Wilk since the data of this research is less than 50.

In order to determine whether the two data groups have normal or abnormal distribution, the writer would examine the *p*-value. The *p*-value would determine whether the hypothesis was accepted or rejected. Below are the statements of hypotheses that the writer would use in drawing the conclusion about the normality of the data.

$H_0$ : The data in the group of assessment is distributed normally.
$H_A$ : The data in the group of assessment is not distributed normally.

The following figure is the result of the normality test. As stated above, the writer would focus on the column of Shapiro-Wilk test and the *p*-value (or the value of *Sig.*) in the said column.

Based on the information above, it can be seen that the *p*-value for the average values of the teacher/lecturer assessment is 0.036 whereas the *p*-value for the average values of the peer assessment is 0.045. Both *p*-values are less than 0.05. This means that the null hypothesis is rejected and the alternative hypothesis is accepted. In other words, the data in both teacher/lecturer assessment and peer assessment are not normally distributed.

Since the result of normality test showed abnormal distribution for both data groups, the writer used Wilcoxon Signed-Rank test to find out whether or not the teacher/lecturer assessment and peer assessment have significant differences, as explained in the data analysis procedures. In order to determine whether there are significant differences or not, the *p*-value needed to be examined further. The hypotheses for this analysis can be stated as follows:

$H_0$ : There is no significant difference between the average values of lecturer assessment and peer assessment.
$H_A$ : There is a significant difference between the average values of lecturer assessment and peer assessment.

Below is the result generated from Wilcoxon Signed-Rank test in SPSS.

**Table 3. Result of Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Average (Lecturer) | .197 | 15 | .121 | .872 | 15 | .036 |
| Average (Peer) | .183 | 15 | .187 | .879 | 15 | .045 |

a. Lilliefors Significance Correction

**Table 4. Result of Wilcoxon Signed-Rank Test**

**Test Statistics[b]**

| | Average (Peer) - Average (Lecturer) |
|---|---|
| Z | -1.308[a] |
| Asymp. Sig. (2-tailed) | .191 |

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Based on the figure above, it can be seen that the *p*-value is 0.191, which is greater than 0.05. This means that the null hypothesis is accepted and that there is no significant difference between the average values of lecturer assessment and peer assessment.

The findings generated above showed that the assessments conducted by teacher/lecturer and students' peers did not have great differences. Both assessments certainly had differences in the average values. However, the differences were not very distinctive. This also meant that peer assessment could be considered as reliable as the teacher/lecturer assessment. The findings of this research were in line with the results generated from the studies by Burke (1969, as cited in Freeman, 1995), Freeman (1995), and Topping (1998), which stated that there was no significant differences between peer and teacher assessments.

There were two possible causes of the similarity between the results generated from peer assessment and teacher/lecturer assessment. The first probable reason was due to the same framework used in the assessment. Both lecturer and students used the same scoring rubric by Jacobs et al. (1981), which specifies five areas of evaluation: Content, Organization, Vocabulary, Language Use, and Mechanics (see Appendix). The lecturer also taught the students the steps that they needed to follow and the points that they need to pay attention to. Moreover, the lecturer reminded the students to give not only scores but also elaborated explanation why they gave those scores and how their peers could improve their writings further. In brief, the students were trained so that they would closely follow the agreed benchmark, rules and regulation, and methods during the assessment process. This certainly limited the students' freedom and restricted their creativity. However, at the same time, the students learned to put their subjectivity aside and to evaluate their peers' writings as objectively as possible. This result and reasoning is consistent with the result of research conducted by Ramon-Casas, Nuño, Pons, and

Cunillera (2018) which stated that the use of structured set of standards could help in promoting the validity and reliability of the result of peer assessment.

The second probable reason for the similarity is due to the nature of scoring rubric. As mentioned above, the rubric by Jacobs et al. (1981) specifies five areas of evaluation: Content, Organization, Vocabulary, Language Use, and Mechanics (see Appendix). Each area of evaluation has its own range of scoring. The maximum point for Content area is 30 points, Organization 20 points, Vocabulary 20 points, Language Use 25 points, and Mechanics 5 points. Jacobs et al. (1981) provided description for each range of scores, and the range of scores is not relatively large. For example, for "very good" to "excellent" content of writing, the range of score is between 27 to 30 points. If the lecturer and the students had similar perception about excellent or very good content, they could give the score of 27, 28, 29, or 30 for the Content area. Even though the lecturer and the students might give different score for the excellent content, the range of difference would be very small and would not make a big difference, which was probably one of the reasons why there was no significant difference between the average results from teacher/lecturer and peer assessments.

The pedagogical implication derived from the result of this research is that peer assessment can be utilized and implemented in writing classes. The high agreement between the results of peer assessment and teacher/lecturer assessment also shows that it is highly probable that peer assessment can be a substitute for teacher/lecturer assessment as long as sufficient scaffolding and proper standards are prepared and set beforehand.

## E. CONCLUSION

This research aims to discover the reliability of peer assessment and the students' perception on peer assessment. The term 'reliability' in this case can be defined as the level of agreement between the lecturer and the students when giving scores. When the result of peer assessment and teacher/lecturer assessment are compared, there should not be any significant differences between those two forms of assessment. If the result showed no significant differences, it meant that the level of agreement between the scores generated from peer assessment and teacher/lecturer assessment was quite high, and thus, reliable. On the other hand, if the calculation and hypothesis testing showed there were significant differences, it could be concluded that the level of agreement was low, and thus, unreliable.

As seen in Findings and Discussion section, based on the result of normality test, the writer had decided to utilize Wilcoxon Signed-Rank test. After the calculation and the hypothesis testing, it was found that the average scores generated from peer assessment and teacher/lecturer assessment did not differ significantly. Therefore, it could be concluded that peer assessment was reliable in this study. The writer suspected that the reasons for the slight differences were because of the usage of the same scoring rubric, which put the assessment into certain perspectives, and the range of scoring in the rubric, which limits the scoring margin. The conclusion drawn in this study was in line with the results of the studies by Burke (1969, as cited in Freeman, 1995), Freeman (1995), and Topping (1998). All three previous studies mentioned that peer assessment has high level of agreement with the teacher assessment, and this study had also proven this point. This means that peer assessment can be considered as one of the options for alternative assessment. Through proper and thorough scaffolding and instructions, the students might be able to substitute for teachers or lecturers in terms of assessing writing tasks.

For future researchers who are interested in conducting further studies on this topic,

it is suggested to do it on a larger scope. This research only utilized small number of samples. Research with larger samples might generate different and interesting findings that could be compared to the results of this study. Another suggestion is to study the students' perceptions on peer assessment. The writer could not analyze this aspect because the writer was also the lecturer who taught the class. Any attempts to elicit the students' perceptions and opinions on the practice of peer assessment could result in bias since the students might feel compelled to answer in positive light. This aspect, nevertheless, can be studied by future researchers and may enrich the current literature.

The last suggestion for the future researchers is to study the effect of feedbacks given during the process of peer assessment. The effect of peer feedbacks has been a largely-discussed topic in the literature. The future researchers could observe the scores after the students are given the peer feedbacks and see whether there has been an improvement on their writing skills.

## REFERENCES

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York, NY: Pearson Education.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93–121. https://doi.org/10.1191/0265532205lt298oa

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: A step-by-step approach*. New Jersey, NJ: John Wiley & Sons.

Dewi, H. D. (2015). *Comparing Two Translation Assessment Models: Correlating Student Revisions and Perspectives*. Kent State University. Retrieved from https://etd.ohiolink.edu/pg_10?::NO:10:P10_ETD_SUBID:109815

Everhard, C. J. (2015). Investigating Peer- and Self-Assessment of Oral Skills as Stepping-Stones to Autonomy in EFL Higher Education. In C. J. Everhard & L. Murphy (Eds.), *Assessment and Autonomy in Language Learning* (pp. 114–142). New York, NY: Palgrave Macmillan.

Freeman, M. (1995). Peer Assessment by Groups of Group Work. *Assessment & Evaluation in Higher Education*, *20*(3), 289–300. https://doi.org/10.1080/0260293950200305

Harris, D. P. (1969). *Testing English as a second language*. New Delhi, ND: Tata McGraw-Hill.

IBM SPSS Statistics for Windows. (n.d.).

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science*, *39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6

Kubiszyn, T., & Borich, G. D. (2013). *Educational testing & measurement: Classroom application and practice* (10th ed.). New Jersey, NJ: John Wiley & Sons.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Purnawan, A. (2015). Peer Assessment as the Main Methd for Assessing Students' Writing: A proto-Design for Developing EFL Lesson Plans. In *The 62nd TEFLIN International Conference 2015* (pp. 260–267).

Ramon-Casas, M., Nuño, N., Pons, F., & Cunillera, T. (2018). The different impact of a structured peer-assessment task in relation to university undergraduates' initial writing skills. *Assessment and Evaluation in Higher Education*, *44*(5), 653–663. https://doi.org/10.1080/02602938.2018.1525337

Roberts, T. S. (2006). *Self, Peer and Group Assessment in E-Learning*. Hershey, PA: Information Science Publishing.

Rofiudin. (2011). Improving Students ' Ability to Write Descriptive Texts through Peer Assessment. In *The 58th TEFLIN International Conference 2011*. Semarang: IKIP PGRI.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998). Creating a Learning Environment by Using Self-, Peer- and Co-Assessment. *Learning Environments Research*, *1*(June 2014), 293–319. https://doi.org/10.1023/A

Sumekto, D. R. (2014). Higher Education Students' Perception about Peer Assessment Practice. In *The 61st TEFLIN International Conference 2014* (pp. 1137–1141). Solo: Universitas Negeri Solo.

Thode, H. C. (2002). *Testing for normality*. New York, NY: Marcel-Dekker, Inc.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Tsui, A. B. M., & Ng, M. (2000). Do Secondary L2 Writers Benefit from Peer Comments? *Journal of Second Language Writing*, *9*(2), 147–170. https://doi.org/10.1016/S1060-3743(00)00022-9

Weimer, M. (2013). *Learner-Centered Teaching: Five Key Changes to Practice* (2nd ed.). San Francisco, California: Jossey-Bass.

White, E. (2009). Student Perspectives of Peer Assessment for Learning in a Public Speaking Course. *Asian EFL Journal - Professional Teaching Articles*, *33*, 1–36.

**Appendix.** Rubric for Peer and Teacher/Lecturer Assessment (Jacobs et al., 1981)

**ESL COMPOSITION PROFILE**

STUDENT DATE TOPIC

| | SCORE | LEVEL | CRITERIA | COMMENTS |
|---|---|---|---|---|
| CONTENT | | 30-27 | EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic | |
| | | 26-22 | GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail | |
| | | 21-17 | FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic | |
| | | 16-13 | VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate | |
| ORGANIZATION | | 20-18 | EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive | |
| | | 17-14 | GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing | |
| | | 13-10 | FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development | |
| | | 9-7 | VERY POOR: does not communicate • no organization • OR not enough to evaluate | |
| VOCABULARY | | 20-18 | EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register | |
| | | 17-14 | GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured* | |
| | | 13-10 | FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured* | |
| | | 9-7 | VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate | |
| LANGUAGE USE | | 25-22 | EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions | |
| | | 21-18 | GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured* | |
| | | 17-11 | FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured* | |
| | | 10-5 | VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate | |
| MECHANICS | | 5 | EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing | |
| | | 4 | GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured* | |
| | | 3 | FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured* | |
| | | 2 | VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate | |

TOTAL SCORE READER COMMENTS