Journal of English Language and Education
ISSN 2597- 6850 (Online), 2502-4132 (Print)
Journal Homepage: https://jele.or.id/index.php/jele/index

OPEN ACCESS

Article

Can Peer Assessment Be Reliable as an Evaluating Instrument?

https://doi.org/10.31004/jele.v10i4.1255

* Rumondang Miranda Marsaulina[1]
[a]Institut Teknologi Del, Indonesia
Corresponding Author: mirandarumondang@gmail.com

ABSTRACT

Discourses asserting peer assessment on English learning process particularly regarding its either effectiveness or downfalls as an alternative evaluation method for English as a Foreign Language (EFL) learners have been disseminated. However, doubts on the lack of peer assessment's capacity as an evaluating tool remains to need more validation by a study examining its reliability in a wider learning context to ensure if the method could be as reliable as teacher's grading leading to a theory that peer assessment can serve for reducing teacher's load especially for big classes. In that connection, this study aims to examine the reliability of peer assessment for a big class of first-year non-native English speaking university students majoring in software engineering but already passing English grammar and vocabulary for composing short text genres in their earlier semester. Methods used for collecting and analyzing the data were Wilcoxon reliability and Bivariate Pearson Correlation tests to compare students' peer assessment and lecturer grading on narrative texts written by 56 software engineering students. The finding shows peer assessment as a tool for evaluating students' writing quality has been in low reliability indicated from the incompatibility between the students' peer assessment quality and the lecturer's grading result. This study contributes to present evidence that peer assessment should be out of consideration as an instrument for evaluating the writing produced by non-native English speaking students despite their passing subjects expected to have enabled them to compose a narrative writing. The conclusion is peer assessment is weak in effect on relieving teacher's assessment load in a big writing class of English for Foreign Learning (EFL) students in spite of their English grammar and vocabulary acquisition at a certain level, though the method might serve for giving non-grading related advantages such as promoting students' metacognition.

Keywords: *Lecturer Grading, Peer Assessment, Reliability, Results*

Article History:
Received 15th July 2025
Accepted 11th August 2025
Published 21st August 2025

INTRODUCTION

Peer assessment have been theorized as one's effort to review the amount, level, value, worth, quality, or success of his or her peer's written work, oral presentations, portfolios, test performance, or other skilled behaviours (Topping, 2009, as cited in Yin et al., 2022). Studies on peer assessment have been dominated with its positive impacts and multiple benefits (Yin et al., 2022; Li et al., 2020) identifying across a wide range of subject areas, education levels, and assessment types when adapted to a specific classroom context (Double et al., 2020).
In that connection, works on peer assessment have been more likely to revolve on its function that can boost students' metacognitive skills in which students are trained to take more responsibility for their learning and enhancing learning outcomes by conducting peer assessment (Jongsma et al., 2023; Widyawati, 2018).

Meanwhile, studies suggesting the reliability of peer assessment as an evaluation instrument were also reported. Comer et al. (2014) reported peer assessment could help assess students' tasks in large volumes. Zhang et al. (2020) inquired peer assessment provided scoring-based evaluation for students' learning achievements. Gupta et al. (2019) and Halim (2021) similarly found peer assessment helped decrease teacher's assessing load for medium sized classes comprising 25 to 30 students. Meletiadou and Tsagari (2014) running the Pearson

Correlation test in an investigation on the reliability and validity of peer assessment of learning writing in a secondary school in Cyprus reported a very high correlation between the teacher's and 40 carefully trained and guided non-native English students' marks on their peer's writing. At the tertiary education level, Wagner (2016) conducting a case study on the result of peer assessment on an 2000-2500 word essay written by final year university students in international business management argued that only few students complained about the grades they received from their peers. Instead, this peer assessment was observed to have impacted positively on the students' learning development, preparing them better for the final exam. Salehi and Sayyar (2017) analyzing the reliability and validity of peer assessment on three-paragraph essays of 32 upper-intermediate Iranian English learners pointed out the method was reliable and valid for written production tasks due to the high correlation between the grades given through peer assessment and those of teacher scoring on students' writing. Likewise, Halim (2021) employing the quantitative method to examine the reliability of peer assessment carried out by 15 university students of an English intermediate writing class was led to a finding that the scores produced by peer assessment were relatively similar to those of the teacher, arising the conclusion peer assessment was reliable as a form of alternative assessment in writing classes. In her qualitative research synthesizing a book chapter and 23 peer-reviewed articles, Damanik (2022) emphasizes peer feedback or peer assessment will increase adult students' learning engagement and collaborative learning skills resulting in enhancing their English writing skills through fostering critical thinking and facilitating meaning negotiations if the students have been trained to conduct the assessment methodically.

Reasonings on the reliability of peer assessment as an instrument for evaluating students' learning results have been attempted to counter by some researchers. Adachi et al. (2018) and Zhang et al. (2020) similarly contended teachers' doubt of its reliability. In their works, (Reddy et al., 2021; Wilson et al., 2015; Yucel et al., 2014) noted those disagreeing with the reliability of peer assessment highlighted its unfairness. This subjectivity-related argument was also supported in the enquiry performed by (Nicol et al., 2014; Yucel et al., 2014; Poverjuc et al., 2012) exposing peer assessment was likely to cause students' resistance and dissatisfaction due to their negative perspectives for the quality of peer assessment compared to that of teacher. Fleckney and Vaz-Serra (2024) conducting a systematic synthesis of 116 papers on how to design effective peer assessment processes discovered strong evidence that peer assessment would be only most effective if it was limited to a form of formative peer feedback in which students, instead of marking their peers' works, merely provided comments expected to help modify their peers' thinking and behaviour to improve their writing.

However, almost all of the academic efforts contesting positive findings on peer assessment for evaluating students' learning process and results only elaborate negative perspectives or opinions either from teachers' side or students' on the integrity of peer assessment without a concrete proof that peer assessment should be indeed excluded from assessment strategies due to its questioning reliability. So far, only Fleckney and Vaz-Serra have literally demonstrated peer assessment should be prevented from being employed in grading students' learning results. Scarce studies actually deducting the non-reliability of peer assessment as a method for judging students' learning accomplishment have been also weakened by limited learning contexts targeted in the studies hypothesizing the non-reliability of peer assessment. For instance, none of previous scientific attempts scrutinize how peer assessment has resulted in the foreign language writing performance of first year non-native English speaking tertiary level students already having knowledge of the grammar and vocabulary for composing some text genres including narrative genre. Therefore, having observed the shortage of analyses exploring the non-reliability of peer assessment as an assessment form in a broader and varied learning context, a new research seeking for a more consolidated proposition whether peer assessment as a student-centered method impacts students' learning evaluation productively remains useful and important. The question derived from the research gap inquired in this study was how reliable peer assessment was actually as an alternative instrument for evaluating the skills of first year non-native English speaking software engineering students already passing fundamental English grammar and vocabulary expected to enough equip them in writing an English narrative passage. The objective of this research was to examine the reliability of peer assessment on a 250 word narrative writing task completed by 56 first-year software engineering students independently. This research finds its significance in an insight production into peer assessment confirming the non-reliability of peer assessment counting on students' English writing competence as an assessment tool equal to teacher's grading for complex and time consuming assessment tasks, such as writing tasks.

METHOD

This study was carried out by employing the quantitative method. This method was selected because of the research objective analysing the comparisons and correlations between the results of peer-to-peer grading and the grades provided by the lecturer as the instructor/teacher to examine the reliability of peer assessment when evaluating students' skills of choosing proper vocabulary, using grammar and syntax and organizing coherent sentences to produce a narrative paragraph.

Furthermore, non-parametric procedure Wilcoxon signed rank test was applied for analyzing the strength of association between peer assessment's results and lecturer grading's scores. Wilcoxon test was selected as the nature of data in this study fulfilled most assumptions that had to be passed (Laerd Statistics, 2018). First, each subject (each student's narrative writing task) in this study was measured on two occasions (through peer assessment and lecturer grading) on the same dependent variable (student's narrative writing performance/skills). The related groups or matched pairs (the same subjects present in both occasions) condition occurring in this study met one of the Wilcoxon signed-rank's assumptions making the test selected to compare these same subjects in this "matched-pairs" study design. The other assumption required by Wilcoxon test passed in this study was the dependent variables measured were ordinal.

Participants

The participants were 56 first-year Software Engineering students passing English fundamental grammar and daily vocabulary in various contexts of writing and speaking in their earlier semester. Based on their average grade class for integrated English skills course one, the students' grammar and vocabulary skills could be categorized as lower intermediate level expected to be competent in producing a short narrative paragraph on a free topic as assigned in this study. This grammar and vocabulary level was expected to provide a strong knowledge for students in assessing their peer's narrative writing quality which should give the results correlated to those of the lecturer's grading.

Instruments

The data were primary, derived from the students' writing task scores resulted from peer assessment and the lecturer (instructor)'s grades on 250 word narrative paragraphs of 56 students. The research instruments were students' narrative passages and a holistic writing rubric comprising three fundamental writing features reflecting one's writing skill which were appropriate uses of vocabulary, grammar along with syntax, and text organization. This rubric was used by both peer students and the lecturer to mark the students' narrative writings.

Procedures

The research procedure consisted of three stages. The first stage was implemented in the class for 75 minutes. Each participant was required to write a 250 narrative paragraph on a free topic for 50 minutes in the classroom. They were expected to incorporate all fundamental grammar and daily vocabulary they had learned in the paragraph. While writing, they were prohibited to open any kinds of dictionaries and lecture notes. After finishing their paragraphs, they had to turn in their work each to the lecturer, who moreover distributed those writings randomly among the class for peer-to-peer assessment. Before students began assessing their peer's narrative writing, they were given a clear instruction for grading using the shared writing rubric. Each student was asked to grade their peer's skills in using vocabulary, applying grammar in context, and organizing the coherent text based on the narrative text quality. The peer assessment was set to take place for 25 minutes, and students furthermore submitted the peer assessment results (grades and feedback comments) to the lecturer. The second stage of this research included the lecturer's sessions of grading the same 56 narrative paragraphs outside classrooms. It took the lecturer three weeks to mark all students' writings. The third stage of the research implemented for 5 weeks involved tabulating the assessment results of target writing components made by students and the lecturer herself and running a number of statistical tests that were fit to analyze and verify the data for yielding the interpretations used to answer the research question and meet the objective of study.

Data Analysis and Verification

In this matched-pairs study, before the reliability of students' writing scores generated through peer assessment was examined, students' scores graded by peer assessment and by teacher supplying the data for this study were tested utilizing Kolmogorov Smirnov and Shapiro-Wilk normality tests to determine the typical data distribution on students' each writing component consisting of vocabulary use, grammar and syntax, and text organization. Based on the normality tests, all data were found to be unusually distributed. Since all data were abnormally distributed, the comparisons of median scores between two assessments for evaluating students' skill for each writing component were analyzed by employing non-parametric procedure Wilcoxon signed rank test to ensure the reliability of students' grades obtained by peer assessment. Two hypotheses denoted as H_0 and H_a were generated to interpret the results of Wilcoxon reliability test:

H_0 : "There is no significant difference between the median score of peer assessment and that of lecturer grading on students' vocabulary use, grammar and syntax, and text organization" indicating peer assessment is as reliable as lecturer grading

Ha : "There is a significant difference between the median score of peer assessment and that of lecturer grading on students' vocabulary use, grammar and syntax, and text organization" indicating peer assessment is unreliable as an alternative form of lecturer grading

Then, the test results were interpreted according to the explanation made by Raharjo (2025, para. 13): (1) if Asymp. Sig. (2-tailed) value < 0.05, H_o is accepted, Ha is rejected. (2) if Asymp. Sig. (2-tailed) value > 0.05, H_0 is rejected, Ha is accepted

These reliability tests' results were verified using the Bivariate Pearson correlation test. This test is useful to determine the linkage strength and the correlation degree between peer assessment and lecturer grading. The Pearson's correlation coefficient or *r* value becomes the basis for interpreting the results of this Bivariate Pearson correlation test (Intellectus Consulting, n.d.) elaborated as follows: (1) a positive *r* value expresses a positive relationship between two variables. (2) a negative *r* value indicates a negative relationship between two variables. (3) a zero *r* value indicates no relationship between the variables at all. (4) degrees of correlation: (a) perfect : *r* value near ±1. (b) high degree/strong correlation : *r* value between ±0.50 and ±1. (c) moderate degree/ moderate correlation: *r* value between ±0.30 and ±0.49. (d) low degree/ weak correlation: *r* value below +0.29. (e) no correlation : *r* value = 0

FINDINGS AND DISCUSSION

Wilcoxon reliability test was employed for each writing component that indicates students' skills of applying suitable vocabulary, using grammar, and arranging text coherently in their narrative writing. These reliability test's results were further verified using the Bivariate Pearson correlation test. Before Wilcoxon reliability test was applied for the data, the data based on students' narrative writing grades obtained through peer assessment and lecturer's grading were tabulated and tested using Kolmogorov-Smirnov and Shapiro-Wilk normality tests to determine data distribution types. Table 1 below compares the raw data originated from students' scores of their narrative writing resulted from peer-to-peer grading and lecturer's marking.

Table 1. Comparisons Between the Results of Peer Assessment and Lecturer Grading on Software Engineering Students' Narrative Writing

Student	Vocabulary		Grammar & Syntax		Organization	
	Peer assessment	Lecturer's grading	Peer assessment	Lecturer's grading	Peer assessment	Lecturer's grading
1	9	8	13	11	5	5
2	10	10	8	15	4	5
3	9	7	14	12	4	4
4	8.5	10	9	15	5	5
5	7	10	11	14	3	4
6	9	9	10	11	3	3
7	8.5	10	14	14	5	5
8	9	9	15	15	5	5
9	8	10	10	14	4	4
10	9	10	12	15	4	5
11	7	10	13	15	4	5
12	10	8	11	8	4	4
13	9	7	13	10	5	3
14	7	7	10	7	4	3
15	6	7	11	7	3	3
16	8	6	10	8	5	3
17	8	9	9	11	5	5
18	8	9	9	9	5	4
19	10	10	10	8	5	3
20	9	8	11	8	3	3
21	8	6	9	6	4	3
22	8	7	12	10	3	3
23	7	9	10	14	4	5
24	9	7	10	7	3	3
25	7	8	12	12	4	4
26	9	8	13	12	5	4
27	11	10	9	8	5	3
28	3	8	4	10	3	3
29	8	7	10	7	4	3
30	8	7	9	8	4	3
31	8	7	8	7	4	3
32	8	8	9	8	4	4
33	8	9	10	13	5	2
34	10	10	13	13	4	4
35	8	9	13	14	3	4
36	10	10	10	14	5	4
37	9	6	9	12	4	4
38	8	9	10	13	4	3
39	8.5	7	13	8	5	3
40	8	7	9	8	3	3
41	8	6	12	8	4	3
42	8	7	12	12	4	3

43	9	8	12	10	4	4
44	9	9	12	10	3	4
45	10	8	10	10	3	4
46	7	6	7	6	5	4
47	9	7	12	12	5	3
48	9	9	12	12	4	3
45	10	8	10	10	3	4
46	7	6	7	6	5	4
47	8	5	12	6	3	4
48	8	8	12	12	4	4
49	8	9	12	11	5	5
50	9	8	12	11	4	4
51	9	8	12	10	3	3
52	10	10	11	9	3	3
53	10	7	5	3	5	3
54	0	8	0	8	0	8
55	9	9	5	4	5	4
56	9	8	4	3	4	3

Once the results of peer assessment and those of the lecturer grading were gathered and tabulated, the normality of data distribution for each writing component was analyzed first to determine which statistics test would fit to calculate the reliability of peer assessment as a whole. Below is the table that displays the results of Kolmogorov-Smirnov and Shapiro-Wilk tests determining the normality of data distribution for peer assessment's scores and lecturer's marks on students' vocabulary use in their narrative text.

Table 2. Normality Tests on the Results of Peer Assessment and Lecturer Grading on Software Engineering Students' Vocabulary Use in Their Narrative Writing

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Peer assessment	.212	56	.000	.867	56	.000
Lecturer's grading	.150	56	.003	.916	56	.001

a. Lilliefors Significance Correction

As seen from table 2, both Kolmogorov Smirnov and Shapiro-Wilk tests results show sig. p-values < 0.05. The sig. p-value of Kolmogorov-Smirnov test is 0.00 for peer assessment and 0.003 for lecturer grading, whereas the sig. p-value of the Shapiro-Wilk test is 0.000 for peer assessment and 0.001 for lecturer grading. These sig. p-values represent the scores generated through peer assessment and the lecturer grading were unusually distributed. Due to this abnormal data distribution, the reliability of peer assessment as an alternative method for evaluating students' skills in choosing suitable vocabulary for composing sentences in their narrative text was tested using Wilcoxon reliability (signed ranks) test as a non-parametric procedure in place of paired-samples t-test applicable for normal distribution only (Corder & Foreman, 2009).

Table 3. Wilcoxon Reliability Test On The Results Of Peer Assessment And Lecturer Grading On Software Engineering Students' Vocabulary Use In Their Narrative Writing

		Ranks		
	Statistics	N	Mean Rank	Sum of Ranks
Lecturer's grading-Peer assessment	Negative Ranks	28[a]	22.55	631.50
	Positive Ranks	16[b]	22.41	359.50
	Ties	12[c]		
	Total	56		

a. Lecturer's grading < Peer assessment

b. Lecturer's grading > Peer assessment

c. Lecturer's grading = Peer assessment

Test Statistics[a]

	Lecturer grading-Peer assessment
Z	-1.629[b]
Asymp. Sig. (2-tailed)	.103

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks

Table 3 shows the Wilcoxon reliability test on the results of peer assessment yields the Asymp.Sig. (2-tailed) value of 0.103 > the significance level of 0.05. This value interpreted as H_0 rejection and Ha acceptance refers to *significant variations between the median score of peer assessment and that of the lecturer grading* on students' vocabulary use for their narrative writing. Such variations indicate that peer assessment is unreliable to evaluate students' performance of using appropriate vocabulary in an English writing.

Table 4. Pearson Correlation Test Between Peer Assessment And Lecturer Grading On Software Engineering Students' Vocabulary Use In Their Narrative Writing

		Peer assessment	Lecturer grading
Peer assessment	Pearson Correlation	1	.284*
	Sig. (2-tailed)		.034
	Sum of Squares and Cross-products	84.710	26.688
	Covariance	1.540	.485
	N	56	56
Lecturer's grading	Pearson Correlation	.284*	1
	Sig. (2-tailed)	.034	
	Sum of Squares and Cross-products	26.688	104.125
	Covariance	.485	1.893
	N	56	56

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

As described in Table 4, the correlation coefficient (*r value*) resulted from Bivariate Pearson test of 0.284 is below +0.29 representing *the weak correlation* between the scores generated through peer assessment and those resulted from lecturer grading, indicating the non reliability of peer assessment for measuring students' vocabulary skill when developing an English passage.

Table 5. Normality Tests On Peer Assessment And Lecturer Grading On Software Engineering Students'grammar Application And Syntax In Their Narrative Writing

	Kolmogrov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Peer assessment	.162	56	.001	.947	56	.016
Lecturer's grading	.152	56	.003	.939	56	.007

a. Lilliefors Significance Correction

Furthermore, on grammar and syntax feature, Kolmogorov Smirnov and Shapiro-Wilk normality tests reveal sig. p-values < 0.05 as well. Table 5 views the results of Kolmogorov-Smirnov test producing sig. p-value at 0.001 for peer assessment and sig. p-value at 0.003 for lecturer grading, while Shapiro-Wilk reveals sig. p-value of 0.016 for peer assessment and sig. p-value of 0.007 for lecturer grading. These values mean abnormal distribution of the data drawn from the results of peer assessment and lecturer grading on students' grammar and syntax in the same writing task. In that connection, Wilcoxon signed-rank test as a non-parametric procedure equivalent to paired-samples t-test was opted again to analyze if peer assessment would be reliable for measuring students' skill in applying correct grammar and syntax in their writing context appropriately.

Table 6. Wilcoxon Reliability Test On The Results Of Peer Assessment And Lecturer Grading On Software Engineering Students' Grammar Application And Syntax In Their Narrative Writing

Wilcoxon Signed Ranks Test

Ranks

	Statistics	N	Mean Rank	Sum of Ranks
Lecturer's grading-Peer assessment	Negative Ranks	32[a]	20.64	660.50
	Positive Ranks	16[b]	32.22	515.50
	Ties	8[c]		
	Total	56		

a. Lecturer's grading < Peer assessment b. Lecturer's grading>Peer assessment
c. Lecturer's grading = Peer assessment

Test Statistics[a]

	Lecturer's grading-Peer assessment
Z	.748[b]
Asymp. Sig. (2-tailed)	.454

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks

Table 6 demonstrates Wilcoxon reliability test generates the Asymp.Sig. (2-tailed) value at 0.454 > the significance level of 0.05 signifying that H_0 is rejected, while Ha is accepted. *This value is interpreted as a significant difference between the median score of peer assessment and that of lecturer grading on students' grammar and syntax in their narrative writing*. Similar to its result on using peer assessment for evaluating students' vocabulary skill when building sentences, results of this Wilcoxon test have revealed peer assessment should be reconsidered as an alternative instrument for evaluating one's English writing performance.

Table 7. Pearson Correlation Test Between Peer Assessment And Lecturer Grading On Software Engineering Students' Grammar Application And Syntax In Their Narrative Writing

		Peer assessment	Lecturer's grading
Peer assessment	Pearson Correlation	1	.365**
	Sig. (2-tailed)		.006
	N	56	56
Lecturer's grading	Pearson Correlation	.365**	1
	Sig. (2-tailed)	.006	
	N	56	56

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

Unlike Pearson correlation test verifying the result of Wilcoxon test which discloses the non-reliability of peer assessment on students' vocabulary skill in writing their narrative text, Bivariate Pearson correlation test for students' grammar application and syntax in the task unveils *a positive correlation coefficient or r value of 0.365, slightly higher than ±0.30 indicating a moderate degree/ moderate correlation between the two assessment methods*. It means peer assessment might be considered in grading students' skills for using correct grammar and syntax in a writing task. Despite the moderate correlation, the r value that is very close to the minimum correlation coefficient indicates peer-to-peer grading for grammar and syntax use in a writing still needs to be completed with a collaborative assessment from the expert or a cross assessment from the lecturer, which will be more time consuming and take double efforts than the mere lecturer grading.

Table 8. Normality Tests On Peer Assessment And Lecturer Grading On Software Engineering Students' Skill For Organizing Sentences In Their Narrative Writing

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Peer assessment	.211	56	.000	.806	56	.000
Lecturer's grading	.267	56	.000	.825	56	.000

a. Lilliefors Significance Correction

On text organization, the Kolmogorov Smirnov and Shapiro-Wilk normality tests display sig. p- values of 0.00 < 0.05 for peer assessment and lecturer grading each. Again, these values represent an abnormal data distribution from both methods. Thus, the Wilcoxon reliability test as a non-parametric procedure was run once more to analyze the reliability of peer assessment in measuring students' skill for organizing their narrative text.

Table 9. Wilcoxon Reliability Test On The Results Of Peer Assessment And Lecturer Grading On Software Engineering Students' Text Organization In Their Narrative Writing

Wilcoxon Signed Ranks Test

Ranks

	Statistics	N	Mean Rank	Sum of Ranks
Lecturer's grading-Peer assessment	Negative Ranks	21[a]	17.67	371.00
	Positive Ranks	10[b]	12.50	125.00
	Ties	25[c]		
	Total	56		

a. Lecturer's grading < Peer assessment
b. Lecturer's grading >Peer assessment
c. Lecturer's grading = Peer assessment

Test Statistics[a]

	Lecturer's grading-Peer assessment
Z	-2.558[b]
Asymp. Sig. (2-tailed)	.011

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks

Table 9 describes the Wilcoxon test results in the Asymp.Sig. (2-tailed) value at 0.011 < the significance level of 0.05 signifying that H_o is accepted, whereas Ha is rejected indicating peer assessment might be reliable to assess students' skill of selecting proper cohesive devices/transition markers to compose their sentences coherently.

Table 10. Pearson Correlation Test Between Peer Assessment And Lecturer Grading On Software Engineering Students' Text Organization Skill In Their Narrative Writing

		Peer assessment	Lecturer's grading
Peer assessment	Pearson Correlation	1	.237
	Sig. (2-tailed)		.078
	N	56	56
Lecturer's grading	Pearson Correlation	.237	1
	Sig. (2-tailed)	.078	
	N	56	56

The Bivariate Pearson Correlation test produces *a positive correlation coefficient* or *r value at 0.237* below *+0.29 implying the low degree or weak correlation between the results of peer assessment and lecturer grading*. It means despite the Wilcoxon test result indicating a small possibility of using peer assessment for grading students' skill of composing their narrative text coherently, the Pearson Correlation test confirms peer assessment had better not be considered as a form of alternative assessment of students' learning writing due to vague reliability of the results.

Discussions

Wilcoxon tests analyzing the reliability of peer assessment on software engineering students' English narrative paragraph verified by the results of Bivariate Pearson Correlation test unveil that peer assessment is lack of its reliability as an assessment form in place of expert's or teacher grading particularly in measuring students' skills of using the suitable vocabulary and grammar in context. In other words, the reliability tests' results have clearly confirmed peer assessment should be excluded as an instrument for evaluating students' abilities in diction and use of grammar and syntax, even for simple and brief paragraphs like the writing task focused in this study. Meanwhile, on the text organization component, the Wilcoxon test brings out a thin possibility for students to participate in assessing their peers' skill in organizing their text coherently. Nevertheless, despite this prospect, the weak correlation between the results of peer assessment and those generated by lecturer grading implies the aversion to engaging students in the assessment process on this feature.

The findings presented in this study are in contrast to those unfolded by (Meletiadow & Tsagari, 2014; Wagner, 2016; Salen & Sayyar, 2017; Gupta et al., 2019, Halim, 2021). Nonetheless, this study shares results that confirm teachers' pessimistic views and students' mistrust on peer assessment reliability described by (Zhang et al., 2020; Adachi et al., 2018; Nicol et al., 2014; Yucel et al., 2014; Poverjuc et al., 2012). The Wilcoxon signed ranks test and Bivariate Pearson correlation test results revealing minor possibilities of employing peer assessment for judging students' grammar and syntax as well as text organizing skills in an English writing indicate a similarity to what have been claimed by Fleckney and Vaz-Serra (2024) on peer assessment's limited effectiveness and Double et al. (2020) contention that peer assessment works effectively only if it is tailored to a unique learning context. This study also signals the flaw of peer assessment as a teacher's alternative instrument consequently leading to a prompt for not including the method for evaluating even a large class, unlike what Comer et al. (2014) have asserted.

In terms of the significance, this study presents evidence verifying previous pessimistic views doubting those praising the efficacy of peer assessment for students' learning results. The finding on the vague reliability of peer assessment has dismissed this method as an alternative evaluation for students' learning product. Peer assessment is unproductive for measuring students' work outcome, though those involved in the process have acquired lower-to-intermediate English level supposed to be optimized for their analyzing the quality of their peer's writing and been given clear directions and guidelines before conducting the assessment. It impacts zero on reducing the marking load of lecturer (instructor/teacher).

However, the functions of peer assessment for promoting students' learning process in a broader context unrelated to grading, for instance as a mean of training students to take more responsibility for their learning (Widyawati, 2018) and improving students' skills in team working, critical thinking and negotiating meanings (Damanik, 2022) have not been unmasked in this study. It is due to the researcher's time limitation to conduct a systematic and comprehensive observation during students' process of grading their peer's writing as well as to investigate students' views on the peer assessment they have engaged through direct interviews. Aside from this, quantitative tests applied for this study have not signaled the reasons why peer assessment conducted by students has no correlation to that of the lecturer, for example, the peer's unfairness discussed by (Reddy et al., 2021; Wilson et al., 2015; Yucel et al., 2014). Thus, a wider range of data collection and varied methods of data analysis like qualitative methods to provoke more elaborated explanation for the shortages should be covered in future research related to the topic of peer assessment as an evaluating instrument for students' learning product.

CONCLUSIONS

Having learned many benefits of involving students in the assessment process including as a solution for teacher's overwhelming responsibilities when having to assess large classes, an inquiry of the reliability of peer assessment on a big class for exercising an English narrative writing in which the members are first year non-native English speaking software engineering students already passing English grammar and vocabulary for composing text genres in the earlier semester became substantial to ensure if this method could replace teacher assessment in English writing sessions.Wilcoxon reliability test and Bivariate Pearson correlation test for verifying the results of Wilcoxon reliability test analyzing the comparison and correlation between the results of peer assessment and those of lecturer grading indicate the non-reliability of peer assessment as a form of lecturer's evaluation instrument. It is verified that such unreliability remains to exist although the students engaged in the peer assessment have possessed lower intermediate English level and university education. All in all, this study suggests the lecturer (instructor/teacher) not take peer assessment as a mere task grading method without cross examination by the lecturer herself. Lecturers (teachers/instructors) are required to continue to mostly tackle students' complex tasks such as writing assignments in whole without relying on the results generated through students' peer assessment.

ACKNOWLEDGEMENTS

I'd like to thank the Language Unit of Institut Teknologi Del for the peer review on the manuscript of this research article. Also, I'd like to thank the Head of Diploma IV Software Engineering study program and LPPM of Institut Teknologi Del for their encouraging this study from the preparation through reporting the results of this study as well as the financial support that allow this study to be disseminated in a nationally accredited journal.

REFERENCES

Adachi, C., Tai, J., & Dawson, P. (2018a). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education*, *43*(2), 294–306. https://doi.org/10.1080/02602938.2017.1339775

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

Damanik, J. Y. (2022). Peer feedback to improve Indonesian adult learners' writing skills: A Review. *JET (Journal of English Teaching)*, *8*(1), 49–58. https://doi.org/10.33541/jet.v8i1.3253

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, *32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Fleckney, P., Thompson, J., & Vaz-Serra, P. (2024). Designing effective peer assessment processes in higher education: a systematic review. *Higher Education Research & Development*, 1–16. https://doi.org/10.1080/07294360.2024.2407083

Gupta, S. D., Abdullah, F., Li, G., & Xueshuang, Y. (2019). Peer Assessment in Writing: A Critical Review of Previous Studies. *Journal of Advances in Linguistics*, *10*, 1478–1487. https://doi.org/10.24297/jal.v10i0.7992

Halim, S.W. (2021). Peer Assessment in University Level: A Preliminary Study on the Reliability. *CaLLs, 7(1)*, pp 1-14.

Intellectus Consulting. (n.d.). Pearson's Correlation Coefficient: A Comprehensive Overview. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/

Jongsma, M. V., Scholten, D. J., van Muijlwijk-Koezen, J. E., & Meeter, M. (2023). Online versus offline peer feedback in higher education: A meta-analysis. *Journal of Educational Computing Research*, *61*(2), 329–354. https://doi.org/10.1177/07356331221114181

Laerd Statistics. (2018). *Wilcoxon signed rank test in SPSS statistics - Procedure, output and interpretation of output using a relevant example.* Laerd.com. https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php

Li, H. L., Xiong, Y., Hunte, C. V., Guo, X. Y., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Meletiadou, E. & Tsagari, D. (2014). An Exploration of the Reliability and Validity of Peer Assessment of Writing in Secondary Education. *Language Learning/Teaching - Education*. https://www.degruyter.com/document/doi/10.2478/9788376560915.p14/pdf?srsltid=AfmBOoqA2fhFFl2-d0iYl50OTWThcoIdPXUtm6iXxhoDS_eaNv6br7XM

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Poverjuc, O., Brooks, V., & Wray, D. (2012). Using peer feedback in a master's programme: A multiple case study. *Teaching in Higher Education*, *17*(4), 465–477. https://doi.org/10.1080/13562517.2011.641008

Raharjo, S. (2014-2025). Panduan Lengkap Cara Melakukan Uji Wilcoxon dengan SPSS. SPSS Indonesia: Olah Data Statistik dengan SPSS. https://www.spssindonesia.com/2017/04/cara-uji-wilcoxon-spss.html

Reddy, K., Harland, T., Wass, R., & Wald, N. (2021). Student peer review as a process of knowledge creation through dialogue. *Higher Education Research & Development*, *40*(4), 825–837. https://doi.org/10.1080/07294360.2020.1781797

Salehi, Mohammad & Sayyar, Zahra. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. Southern African Linguistics and Applied Language Studies. 35. 1-15. 10.2989/16073614.2016.1267577.

Wagner, S. M. (2016) "Peer feedback: moving from assessment of learning to assessment for learning", *Journal of Learning Development in Higher Education*. doi: 10.47408/jldhe.v0i0.335.

Widyawati, W. Y. (2018). Peer assessment for improving writing descriptive text of the tenth graders of senior high school setia budhi Semarang. *ETERNAL (English Teaching Journal)*, *7*(2). https://doi.org/10.26877/eternal.v7i2.2168

Wilson, M. J., Diao, M. M., & Huang, L. (2015). 'I'm not here to learn how to mark someone else's stuff': An investigation of an online peer-to-peer review workshop tool. *Assessment & Evaluation in Higher Education*, *40*(1), 15–32. https://doi.org/10.1080/02602938.2014.881980

Yin, S., Chen, F., & Chang, H. (2022). Assessment as Learning: How Does Peer Assessment Function in Students' Learning? *Frontiers in Psychology. 13 (912568)*, 1-14. https://doi.org/10.3389/fpsyg.2022.912568

Yucel, R., Bird, F. L., Young, J., & Blanksby, T. (2014). The road to self-assessment: Exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education*, *39*(8), 971–986. https://doi.org/10.1080/02602938.2014.880400

Zhang, F., Schunn, C., Li, W., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & Evaluation in Higher Education*, *45*(8), 1073–1087. https://doi.org/10.1080/02602938.2020.1724260