Journal of Modern Research in English Language Studies
Print ISSN: 2676-5357
Online ISSN: 2676-5985

Severity Differences Across Proficiency Levels Among Peer-Assessors

Shahla Rasouli[1], Rajab Esfandiari*[2]

[1]MA in TEFL, Imam Khomeini International University, Qazvin, Iran.
shahla.rasouli@gmail.com
[2]Associate Professor, Department of English Language, Faculty of Humanities, Imam Khomeini International University, Qazvin, Iran. *esfandiari@hum.ikiu.ac.ir*

Article Info

Article type:

Research Article

Received:

07/09/2020

Accepted:

20/11/2020

ABSTRACT

Over the past few years, peer-assessment, as an alternative assessment procedure, has drawn the attention of many researchers. In the study, it was attempted to find what kinds of language components peer-assessors attend to when rating their peers' essays and to investigate whether proficiency levels of peer-assessors make a difference in terms of severity and leniency they exercise. Fifty-eight student raters at Imam Khomeini International University in Qazvin rated five essays, using an analytic rating scale. Paper-based test of English as a foreign language (TOEFL) and five-paragraph essays were used to collect the data. FACETS (version 3.68.1) was used to analyze the data. The results of Facets analysis indicated that advanced peer-assessors had more variability in their severity compared to intermediate peer-assessors. Moreover, the majority of peer-assessors were, on average, more severe than lenient. The results also revealed no statistically significant difference between the ratings of intermediate and advanced peer-assessors. The final finding was that task achievement was the most attended assessment criterion, but grammatical range and accuracy was the least attended assessment criterion. The findings suggest peer-assessors do not attach an equal weight to all assessment criteria. The findings of the study may carry implications for the summative assessment of students' abilities.

Keywords: Criterion, Peer-Assessment, Proficiency Level, Severity

Cite this article: Rasouli, S,. & Esfandiari, R. (2022). Severity differences across proficiency levels among peer-assessors. *Journal of Modern Research in English Language Studies, 9*(2), 173-196.

DOI: 10.30479/jmrels.2022.16763.2014

1. Introduction

The last two decades have witnessed a huge alteration in how student performance is assessed. Alternative assessment has been proposed to refer to evaluating L2 learners' performance which allows researchers to come up with a more holistic approach to student assessment (Minzi & Zhang, 2021; Saito, 2008). Alternative assessment creates conditions to support students' learning (Black & Wiliam, 1998; Zhang et al., 2020), and there are different types of alternative assessment procedures including, checklists, journals, logs, self-assessment, teacher assessment, and peer assessment (Brown & Abeywickrama, 2018).

One of the most commonly used alternative assessment procedures for formative assessment, peer assessment is "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (Topping, 2010, p. 62). Based on this definition, peer assessment contains students' judgment and remarks of other peers, using some pre-established criteria (Loddington, 2008; Li et al., 2021).

Peer assessment has several advantages. Initially, it increases the tendency of learning from peers (Gibbs, 1999; McDowell & Sambell, 1999). Secondly, in the case of adult learners, it establishes relationship between teachers and students (Leach et al., 2001). Thirdly, it develops cognitive thinking (Cheng & Warren, 2005; Davis, 2009). Finally, it reduces teachers' workload (Topping, 2009). The advantages notwithstanding, peer-assessors may exercise severity and leniency when rating essays. Matsuno (2009) argued that peer assessors were more lenient compared to self-assessors. By contrast, Esfandiari and Myford (2013) found that peer-assessors were more severe than self-assessors in their ratings. Since some raters are harsher than other raters, it may threaten the reliability of the ratings. In order to eliminate this bias, Elder et al. (2007) suggested providing raters with training sessions to have more reliable ratings in the second language contexts. Comparing the raters' ratings before and after training, Weigle (1998) stated that training helped raters reduce severity and leniency.

It is important to know what criteria are used in peer-assessment. If students do not receive training about how to use criteria, they may have trouble applying the criteria in their ratings (Orsmond et al., 1996). Further, there are no details regarding the quality of the criteria which are used in particular contexts of peer-assessment (Dancer & Dancer, 1992; Cho et al., 2006). Even though the students are taught the criteria, they may find them difficult and avoid using them, or they may be unable to apply the criteria (Orsmond et al., 1996).

The study tries to find what kinds of language components peer-assessors attend to when rating their peers' essays and to investigate whether proficiency levels among peer assessors make a difference in terms of severity and leniency. Examining severity differences among peer-assessors may prove promising. Studies of this type may carry implications for training purposes, graduate writing courses, and concurrent validity. Students can be supplied with diagnostic information on how to reduce more cases of severity and leniency if their ratings are to be used for summative judgments. Longer training periods may be held to instruct students to best use the rating scale criteria and the guidelines about how to rate essays. Students may be provided with rich feedback regarding their ratings so that they will incorporate it in their ratings. Therefore, we formulate the following research questions to provide answers in this study.

1. To what extent can peer-assessors be severe or lenient when assessing the essays of their peers?
2. Does proficiency level make a difference in peer-assessors' rating of EFL essays?
3. What assessment criteria do peer-assessors attend to when rating EFL essays?

2. Literature Review

In this part, the concept of alternative assessment is introduced and discussed. Followed by this conceptual explication includes a detailed discussion on peer-assessment in language assessment. Severity and leniency are next defined, and the summary of the findings of some previous studies is presented. Finally, the criteria used to assess student performance are explained.

2.1. Alternative Assessment

The notion of alternative assessment can be regarded as "an alternative to standardized testing" (Huerta-Macías 1995, p. 8). Alternative assessments have been considered from four perspectives, namely, technological, cultural, political, and postmodern (Hargreaves et al., 2002; Zhan, 2021). Concerning the technological issues, the way of measuring results and achieving implementation are struggles for teachers; from a cultural perspective, developing assessment criteria with students and explaining them reasonably, and emphasis on interaction between beliefs and values are of highest importance; furthermore, political perspective of alternative assessment considers the act of power and its possible supervision rather than allowing people; and, finally postmodern perspective of alternative assessment

concerns the concept of authentic assessment whose meaning remains questionable.

There are positive characteristics for alternative assessment. Students perform, create, produce, or do something; use real world situations; focus on process and products; and are given information about their strengths and weaknesses (Huerta-Macías, 1995; Kolomuç, 2017). Alternative assessments require problem solving and higher level thinking and involve tasks that are worthwhile as instruction activities (Aschbacher, 1991). Herman et al. (1992) identified a different set of characteristic of alternative assessment, stating that "alternative assessments (a) tap into higher level thinking and problem-solving skills; (b) use tasks that represent meaningful instructional activities; and (c) call upon teachers to perform new instructional and assessment roles" (p. 6).

2.2. Peer-Assessment

Peer assessment can be defined as "an arrangement of peers to consider the level, value, worth, quality, or successfulness of the products or outcomes of learning of other of similar statues" (Topping et al., 2000, p. 150). Products can include writing assignment, portfolios, projects, oral presentations, test performance or other skilled behavior (Topping, 2009). Topping argued that peer-assessment can vary in a number of ways, such as the participant constellation which can be the assessor and assessed in pairs or in groups; directionality, for instance, one way or reciprocal; and objectives, that is, the teacher may target cognitive or metacognitive gains, time saving, or other goals. Peer assessment is an approach in which the members of a group decide the extent to which each member deserves an amount of group mark (Goldfinch & Raeside, 1990).

Some theoretical frameworks have been cited in support of peer-assessment, for example, theories of language development and acquisition such as Vygotsky's (1978) scaffolding and zone of proximal development (ZPD) and interactionist theories of second language acquisition such as Long's (1985). According to Vygotsky (1978), the collaborative nature of peer-assessment activities offers chances for learners to be "scaffolded" in learning through interaction with more knowledgeable peers. The advocates of interactionist theories focus on the communicative nature of group work and on the opportunities of peers to negotiate meaning, which promote comprehension and acquisition. Similarly, Mendonca and Johnson (1994) DiGiovanni and Nagaswami (2001) concentrated on the interaction side and believe that students are able to negotiate meaning, to ask for clarification, to give suggestion, and to practice a wide range of language skills.

The significant benefits of peer-assessment have been identified by researchers, teachers, and peers themselves (Brown & Glasner, 1999; Saito, 2009; Topping, 2009). Students concern themselves about producing excellent work; therefore, they realize they may be judged by their peers (Searby & Ewer, 1997). Peer-assessment helps students to improve certain skills in communication, self-evaluation, and self-criticism (Dochy & McDowell, 1997). Peer-assessment is a useful educational strategy for developing learning and has been found to strengthen students' engagement (Bloxhom & West, 2004). Peer-assessment encourages broad interaction in relation to a task; therefore, through this interaction, teachers and students can comprehend each other well (Boud et al., 2001). Peer-assessment can be appropriate for independent learning; also, it requires students to make independent judgment and provide analyses on the perform of their peers (Boud et al., 2001; Brown et al., 1997; Brown & Glasner, 1999; Brown & Knight, 1994; Brown et al., 1995). The practical benefits from peer-assessment are developing problem-solving skills, saving teacher's time, generating understanding of nature and process of assessment, increasing motivation, and making it easier for the student to reject/interact with feedback (Hansen, 2014).

2.3. Severity and Leniency of Peer-Assessment

Severity and leniency effect is the most serious error that a rater can introduce into a rating setting (Cronbach, 1990). Generally speaking, severity refers to being harsh and leniency has to do with being relaxed. However, in rater-mediated assessments, severity and leniency assume specialised meanings. Myford and Wolf (2004) defined rater severity as a "rater's tendency to assign ratings that are, on average, lower than those that other raters assign". By contrast, rater leniency refers to a "rater's tendency to assign ratings that are, on average, higher than those that other raters assign" (p. 94)

Several strategies have been proposed to try to minimise the impact severity and leniency may have on the measurement of ratings. These strategies are summarised in Myford and Wolfe (2003) as follows: Clear definitions of the traits to be rated have to be given; peer-assessors need to be sufficiently trained; and statistical methods should be used to adjust for peer-assessors' leniency or severity.

2.4. Assessment Criteria in Peer-Assessment

To judge their peers' performance, peer assessors need to use some criteria. Sadler (1987) defined a criterion as "a distinguished property or characteristic of anything, by which its quality can be judged or estimated, or

by which a decision or classification may be made" (p. 194). Moreover, Dochy et al. (1999) argued that the development of criteria through active cooperation between teachers and students was the critical factor for peer-assessment. They found that when criteria in peer-assessment are determined in advance in joint collaboration between teachers and students, the result is more satisfactory. The second finding of the study was that the criteria should be defined operationally and students should be familiarised with them.

Boud (1989) used a nominal group process to identify the criteria that students suggested. They involved students in group exercise to find a common set of criteria and use the criteria for judging individual performance in classroom. He found that "students need to be able to assess themselves in situations in which they have only partial knowledge of the criteria to be used by others and when they may not fully accept the criteria which others will apply to them" (p. 22). Orsmond et al. (1996) reported the method which allows peers to rate products against the individual criteria. The results showed that "there was no significant difference between the tutor and peer mark, for the 'self-explanatory' and 'clear purpose' criteria" (p. 244). Even though students were instructed about how to use the criteria, they were unable to recognise them.

3. Method

3.1. Participants

The research included a paper-based TOEFL test and an IELTS scale. Regarding the writing, 58 Iranian EFL students were asked to write a five-paragraph essay. These students were selected because they formed a homogenous group in terms of writing ability to begin our study. The participants consisted of 58 students, 17 male and 41 female BA students majoring in English Language Teaching and English Translation at Imam Khomeini International University in Qazvin, Iran.

The participants were divided into three groups. The students who obtained 70% of total scores of the paper-based TOEFL proficiency test were classified as advanced peer-assessors, those scoring between 46% and 69% as intermediate peer-assessors, and those whose scores were below 45% were grouped as beginning peer-assessors (Phakiti, 2003). In the present study, only intermediate and advanced peer-assessors were used to assess the essays of their peers, so the peer-assessors whose scores were below 45% (in this study below 19) were omitted. Seven beginning peer-assessors were left out.

3.2. Data Collection Methods

Three assessment instruments were used in this study: students' essays, TOEFT test, and IELTS rating scale. A detailed description for these instruments is given below. Fifty-one five-paragraph essays collected from undergraduate (BA) students were used in this study. The students were enrolled in essay writing courses at Imam Khomeini International University in Qazvin, Iran. The students in Essay Writing classes were taught features of a well-written five-paragraph essay such as organisation, content, transitions and coherence. The students in these classes were also taught various patterns of development, including comparison and contrast essays, cause and effect essays, and enumeration essays. After eight meetings, the instructor told his students that they would take the midterm exam the following week. During the exam, students had 40 minutes to write a five-paragraph essay in 250 words at least.

The second instrument used in this study was a paper-based TOEFL test to divide students into two proficiency levels. This test included 40 grammar items, 50 listening comprehension items, and 50 reading comprehension items. The third instrument was IELTS scale (public version) to rate students' essays. This is a 9-band scale, including four criteria to evaluate IELTS essays. The criteria include task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. Descriptors were used to help raters to assign ratings. Following Marefat and Heydari (2016), the present researchers used Content to stand for Task achievement, Organisation for Coherence and Cohesion, Vocabulary for Lexical Resource, and Grammar for Grammatical Range and Accuracy. Therefore, in the present study, they are used interchangeably.

3.4. Procedure

The following steps were used to carry out this study. First, the students were taught features of five-paragraph essays in eight sessions. After these sessions of instruction, they wrote about the topic "What problems do you think parents face when dealing with their children using the internet. How can this problem be solved?" The students were given 40 minutes to write a five-paragraph essay ranging in length from 250 to 300 words. All the students wrote about the same topic in order to control the topic effect.

Peer-assessors were asked to rate five essays of their peers based on the IELTS scale. Students' names were removed from the papers to preserve anonymity. Peer-assessors were asked to attend a 2-hour training session in which they were briefed on rating to familiarise them with rating procedures. They were also asked to leave comments when necessary about various elements and features of the scripts and correct the students' error if necessary. They were supposed to hand in the rated essay within two weeks.

The peer-assessors received feedback on their assessment of the essay after analysis of the data was completed.

3.5. Data Analysis

First, descriptive statistics were used to analyse the assessment criteria peer-assessors attended to. Independent samples *t*-tests were used to test whether levels of proficiency made a difference in attending to those assessment criteria. An independent samples *t*-test was used to ensure two groups of peer-assessors differed on the ratings they awarded to their peers. In order to ensure the proper functioning of rating scales and to analyse the ratings for severity/leniency, FACETS (version 3.68.1., Linacre, 2011) was used. Facets was also used to identify severe and lenient peer-assessors. Average severity and leniency measures were determined for both intermediate and advanced peer-assessors. Total raw scores were computed to assign peer-assessors into intermediate and advanced groups. Percent figures were used to tally the number of times peer-assessors used assessment criteria.

4. Results and Discussion

4.1. Results

4.1.1. Investigation of the First Research Question

The first research question asked the extent to which peer-assessors could be severe or lenient when assessing the essays of their peers. To answer this question, the researchers used the many-facet Rasch measurement. The following paragraphs describe the answer to this research question.

The raw scores, on a 9-point scale, assigned to the essays by the peer-assessors were submitted to FACETS to model the relationship between the three facets of analysis: the peer-assessors (25 intermediate peer-assessors, 26 advanced peer-assessors), the essays (n = 5), and the assessment criteria (4 assessment criteria: task achievement (TA), coherence and cohesion (CC), lexical resources (LR), and grammatical range and accuracy (GRA). This relationship can be expressed as follows: A peer-assessor + an essay + an assessment criterion → a rating.

Figure 1 is the graphical representation of the relationship between the facets of the model. In this figure, which is technically referred to as Vertical Rulers, furthest to the left is the measurement ruler, labeled Measr. The values of this ruler are in logits, ranging from −3 to +3, with zero being the

mean, negative values showing ratings falling below the mean and the positive values displaying ratings positioned above the mean. Then, each of the other columns shows the elements of a facet positioned on the measurement ruler.

The column labelled + peer-assessors represents severity/leniency, ranging from −2 to +2 logits for advanced peer-assessors and -1 to +1 for intermediate peer-assessors. This means that the advanced peer-assessors showed more variability in their severity compared to intermediate peer-assessors. Except for advanced peer-assessors 43 and 46, all the other peer-assessors are between -1 and +1 logits. This implies, generally, peer-assessors showed less variation in their severity. Also, the most lenient peer-assessor was the advanced peer-assessor 43, and the most severe peer-assessor was the advanced peer-assessor 46.

Column three shows proficiency of the peer-assessors. Although both intermediate and advanced peer-assessors roughly fall on the mean, implying they may be neither severe nor lenient, they exercise differing levels of severity, as shown in this section.

The next column is labelled + Essays. The essays above the mean received low ratings while those below the mean received high ratings. As can be seen, essays 1 and 2 received the lowest ratings, essays 4 and 5 average ratings, and essay 3 the highest rating. The column labelled + Assessment criteria represents item difficulty based on the four rating criteria. It can also be seen that assessment criterion 4 was difficult for students to receive high ratings on; by contrast, assessment criterion 1 was easy for students to receive high ratings on. Finally, the last column shows the IELTS 9-point rating scale, ranging from 0 to 9, as the score band.

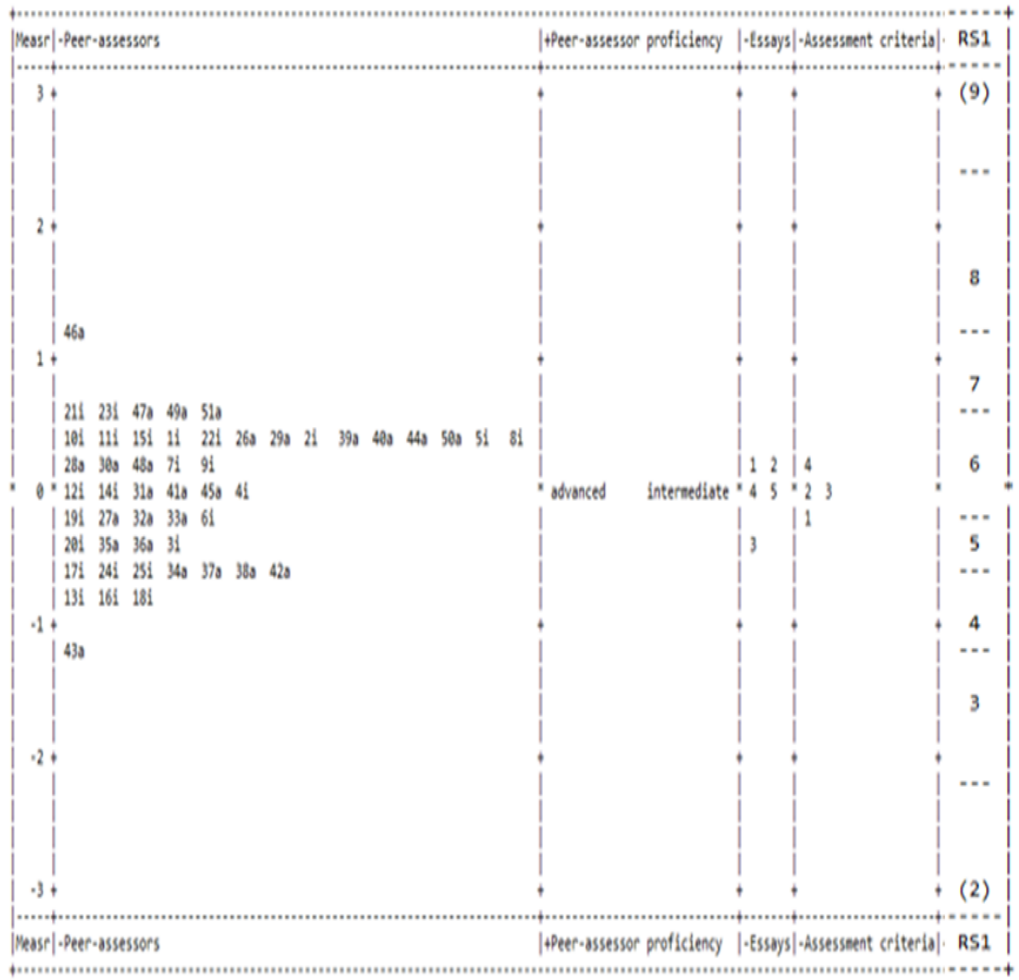
FACETS also produces detailed reports about the performance of individual peer-assessors in terms of total scores and logits (Table 1). It should be noted that total score is the sum of raw scores, on the IELTS 9-point scale, each peer-assessor gave to all the essays for each assessment criterion ($5 \times 4 \times 9 = 180$). The peer-assessors are ordered from the most severe, on top, to the most lenient, at the bottom of Table 1.

As shown in Table 1, advanced peer-assessor 46 was the most severe peer-assessor of all the 51 peer-assessors at +1.23 logits and a total score of 70. Moreover, advanced peer-assessor 43 was the most lenient peer-assessor at −1.28 logits and a total score of 150. More than half of the intermediate peer-assessors (14) were severe, assigning lower ratings to the essays of their peers while 11 intermediate peer-assessors were lenient, assigning higher ratings to the essays of their peers. Like intermediate peer-assessors, more than half of the advanced peer-assessors (15) were severe, but 11 advanced

peer-assessors were lenient. These findings suggest peer-assessors, regardless of their proficiency level, assessed the essays of their peers severely, awarding low ratings to the essays.

Figure 1

Variable Map from FACETS Showing the Relationships between Peer-Assessors, Proficiency of Peer-Assessors, Essays, and Assessment Criteria



Note. i = Intermediate Peer-assessor, a = Advanced Peer-assessor, 1 = Task Achievement, 2 = Coherence and Cohesion, 3 = Lexical Resources, 4 = Grammatical Range and Accuracy

Table 1

Measurement Report for Individual Peer-Assessors

Peer-assessors	Total score	Logit	Error
46a	70	1.23	.18
47a	89	.68	.17
23i	92	.67	.17
21i	93	.64	.17
49a	92	.60	.17
51a	94	.54	.17
2i	72	.42	.19
5i	99	.47	.17
50a	97	.46	.17
22i	100	.44	.17
39a	98	.43	.17
44a	98	.43	.17
11i	101	.41	.17
26a	101	.41	.17
10i	102	.38	.17
15i	103	.35	.17
29a	101	.34	.17
1i	104	.32	.17
8i	104	.32	.17
40a	102	.31	.17
30a	105	.22	.17
9i	109	.17	.18
28a	107	.16	.17
7i	110	.13	.18
48a	108	.13	.18
31a	110	.06	.18
45a	110	.06	.18
4i	114	.01	.18
14i	114	.01	.18
12i	115	-.02	.18
41a	113	-.03	.18
6i	118	-.12	.18
32a	116	-.12	.18
33a	116	-.12	.18
19i	119	-.15	.18
27a	118	-.19	.18
35a	122	-.32	.18
36a	122	-.32	.18
20i	125	-.35	.18
3i	128	-.44	.18
17i	130	-.51	.18
25i	130	-.51	.18
37a	129	-.55	.18
34a	131	-.61	.18
38a	131	-.61	.18
42a	132	-.65	.18
24i	128	-.67	.19
13i	137	-.74	.18
16i	137	-.74	.18
18i	138	-.78	.18
43a	150	-.128	.20

Note. Separation = 2.60, Reliability = .87, Fixed (all same) Chi-square = 376.1, df. = 50, Significance (probability) = .00

The separation index for peer-assessors (N = 51) was 2.60, which suggests that there were about three statistically distinct levels of severity within the peer-assessors. The reliability of the peer-assessors was .87, further confirming the distinct levels of severity among peer-assessors. These severity measures are statistically significant ((χ^2 (50) = 376.1, p < .05).

In addition to individual performance, FACETS generates detailed pieces of information of group performance of peer-assessors. Table 2 compares the overall ratings of the intermediate and advanced peer-assessors. As can be seen, the intermediate peer-assessors have a positive logit value of .01, while the advanced peer-assessors assessed the essays at -0.06 logits. However, this does not mean that intermediate peer-assessors were significantly more severe than advanced peer-assessors in their overall scoring (p = .07). In addition, the Separation Index for the two groups of peer-assessors is 1.48, showing that the variance among the severity of the two groups is about one and a half times the error of estimate. Furthermore, the reliability index of .69 indicates that the analysis is somewhat reliably separating peer-assessors into different levels of severity. Table 2 also shows that both groups have infit and outfit mean square values ranging between .80 and1.20, indicating an acceptable fitness of data (Wright & Linacre, 1994).

Table 2

Overall Measurement Report for Intermediate and Advanced Peer-assessors

Proficiency level of peer-assessors	Total score	Logit	Error	Infit		Outfit	
				MnSq	ZStd	MnSq	ZStd
Intermediate peer-assessors	2930	.01	.03	.86	-.2.3	.87	-2.2
Advanced peer-assessors	2761	-.06	.03	1.14	2.1	1.14	2.2

Note. Separation = 1.48, Reliability = .69, Model, Fixed (all same) chi-square = 3.2 d.f = 1 significance (probability) = .07

The average severity measures for the two groups of peer-assessors, along with their respective standard errors, were as follows: intermediate peer-assessors (0.01 logits, 0.03) and advanced peer-assessors (-0.06 logits, 0.03). The results from the chi-square test of homogeneity indicated that the average severity measures for the proficiency levels of peer-assessors were

all the same, after allowing for measurement error ($\chi^2(1, N = 51) = 3.2$, $p > .05$). An independent *t*-test showed that the average severity measures for the intermediate peer-assessors and advanced peer-assessors were not statistically significant (49) = 1.6493, $p = 0.105$, 95% CI [-0.0153 to 0.1553].

Based on the information in Table 1 and Table 2, although individual peer-assessors showed statistically significant levels of severity differences, no statistically significant differences were found between intermediate peer-assessors and advanced peer-assessors, suggesting the ratings they award, on average, can be used interchangeably.

4.1.2. Investigation of the Second Research Question

The second research question of this study aimed to explore whether proficiency level makes a difference in peer-assessors' rating of EFL essays. Using SPSS software, an independent sample *t*-test was run. Table 3 presents the results.

The significance level for Leven's test is (.715). This is larger than the cut off of .05. Therefore, the assumption of equal variance has not been violated. The sig (2-tailed) value is above .05 (sig = .737). Hence, there was no significant difference on scores for intermediate (M = 112.76, SD=16.36) and advanced (M = 111.23, SD = 15.98; t(49) =.338, p =.737). Therefore, the second research question of the study is confirmed, indicating that there is no statistically significant difference between the scores of intermediate and advanced learners in rating.

Table 3

Independent Samples t-test for Level of Proficiency

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
TRs	Equal variances assumed	.134	.715	.338	49	.737	1.52923	4.52994	-7.57	10.63
	Equal variances not assumed			.337	48.802	.737	1.52923	4.53209	-7.57	10.63

4.1.3. Investigation of the Third Research Question

The third question of this aimed at finding the criteria which peer-assessors attend to when rating EFL essays. The mean of all the ratings for each criterion was calculated to answer this research question. Mean values are shown in Table 4. Mean values are very close to each other. Peer-assessors attended to Task Achievement more than they did to the other criteria. Cohesion and Coherence was the next most attended criterion. Lexical Range was the third most attended criterion for peer-assessors. Grammatical Range and Accuracy was the least attended criterion.

Table 4

Statistics for Criteria

Criteria	TTA	CC	LR	GRA
Mean	28.5882	28.4708	27.8627	27.0588

Four independent samples *t*-tests were used to examine whether intermediate and advanced peer-assessors used each criterion differently. The results are presented below. An independent-samples *t*-test was conducted to compare the Task Achievement scores for intermediate and advanced peer-assessors. As shown in Table 5, there was no statistically significant difference in scores for intermediate students (M = 29, SD = 5.21) and advanced students (M = 27, SD = 5.74; $t(49) = 1.29$, $p = .203$).

Another independent-samples *t*-test was conducted to compare the Coherence and Cohesion scores for intermediate and advanced peer-assessors. As shown in Table 6, there was no statistically significant difference in scores for intermediate students (M = 28.92, SD = 4.79) and advanced students (M = 28.03, SD = 4.44; $t(49) = .682$, $p = .499$).

A third independent-samples *t*-test was conducted to compare the Lexical Resource scores for intermediate and advanced peer-assessors. As shown in Table 7, there was no statistically significant difference in scores for intermediate students (M = 27.56, SD = 4.19) and advanced (M = 28.15, SD = 3.97; $t(49) = -.519$, $p = .606$).

The final independent-samples t-test was conducted to compare the Grammatical Range and Accuracy scores for intermediate and advanced peer-assessors. As shown in Table 8, there was no statistically significant

difference in scores for intermediate students (M = 26.68, SD = 4.25) and advanced groups (M = 27.42, SD = 4.43, t(49) = -.610, p = .545).

Table 5

Independent Samples t-test for TTA

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig.(2-tailed)	Mean difference	Std. Error Difference	95% Confidence Interval of the Difference	
									lower	Upper
TTA	Equal variances assumed	.060	.807	1.291	49	.203	1.98	1.53	-1.10	5.07
	Equal variances not assumed			1.293	48.844	.202	1.98	1.53	-1.09	5.06

Table 6

Independent Samples t-test for TCC

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig.(2-tailed)	Mean difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
TCC	Equal variances assumed	.587	.447	.682	49	.499	.88154	1.29	-1.71	3.48
	Equal variances not assumed			.680	48.343	.499	.88154	1.29	-1.72	3.48

Table 7

Independent Samples *t*-test for TLR

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig.(2-tailed)	Mean difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
TLR	Equal variances assumed	.120	.731	-.519	49	.606	-.59	1.14	-2.89	1.70
	Equal variances not assumed			-.518	48.579	.606	-.59	1.14	-2.89	1.70

Table 8

Independent Samples-t test for TGRA

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	Df	Sig.(2-tailed)	Mean difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
TGRA	Equal variances assumed	.000	.988	-.610	49	.545	-.74	1.21	-3.19	1.70
	Equal variances not assumed			-.610	49	.545	-.74	1.21	-3.19	1.70

4.2. Discussion

The study aimed to determine the extent to which peer-assessors could be severe, or lenient, when assessing the essays of their peers. The study also intended to examine whether proficiency level would make a difference in peer-assessors' rating of EFL essays. Finally, it was attempted to find what assessment criteria peer-assessors would attend to when rating EFL essays.

First, it was found that although individual peer-assessors showed statistically significant levels of severity differences, no statistically significant differences were found between intermediate and advanced peer-assessors, implying that, irrespective of their proficiency levels, peer-assessors assessed the essays of their peers severely awarding low ratings to the essays. Further, according to peer-assessors' severity/leniency logits, for advanced peer-assessors, the ranging was from -2 to +2, and -1 to +1 for intermediate peer-assessors This suggests that the advanced peer-assessors showed more variability in their severity compared to intermediate peer-assessors.

This finding is in line with the findings of some other studies (Esfandiari & Myford, 2013; Nakamura, 2002; Saito & Fujita, 2004, 2009). Generally, these studies showed distinct levels of severity/leniency within peer-assessors. Esfandiari and Myford (2013) found that the average severity measures for the peer-assessors were not statistically significant. Moreover, assessor separation index for the peer-assessors (n = 136) was 3.7, which suggests that there were about three statistically distinct levels of severity within that assessor type. Further, they found that peer-assessors tended to rate significantly more severely. Further, the results of the present study confirm the findings of Hanrahan and Issacs (2001), who showed that peer-assessors were more severe because they were more critical of peers in assessing the essays. Moreover, Nakamura (2002) found that peer-assessors were more severe in rating essays.

This finding, however, is not consistent with that of Brown (1995), who reported that raters with different levels of proficiency differed in perceiving the assessment criteria and applied the criteria differently. Further, in the study of Weigle (1994), unexperienced raters were more strict and inconsistent than experienced raters. Saito and Fujita (2008) indicated that the level of proficiency did not make a difference in peer-assessors' rating. At the same time in the study of Berg (1999), after rater training which made students more proficient than before the training, they did not find any difference in their ratings. Moreover, in the study of Saito and Fujita (2004), who examined the severity and leniency of peer-assessors, they found that peer-assessors were comparatively more lenient than severe.

The findings of quantitative data analysis revealed that the level of proficiency was not an important factor, and there was no statistically

significant difference between the ratings for intermediate and advanced groups. This was one of the major findings of the study which demonstrated that proficiency level did not make a difference in peer-assessors' rating of EFL essays. Lumley (2002) noted that level of proficiency does not necessarily lead to differences in the ratings of peer-assessors, arguing that even after rater training for unexperienced raters, no significant change was observed.

The next finding of this study was that peer-assessors paid the highest attention to Task Achievement and the least attention to Grammatical Range and Accuracy. In other words, they were more concerned with Content and Organization than Grammar and Vocabulary. These findings support the finding of Lee (2009) on Korean raters, who found that raters were more concerned with content and vocabulary than other criteria. Similarly, Kuiken and Vedder (2014) reported that Dutch and Italian raters attached more value to discourse (organisation and content) than surface (grammar and vocabulary) features. These results, however, are not consistent with those of Connor-Linton (1995), who compared the American and Japanese students' ratings and noted that they tended to focus on surface-level features. Similarly, Marefat and Heydari (2016) found that Iranian raters perceived grammar and vocabulary as the most important criteria and content and organization as the least important criteria to rate EFL essays.

This last finding is surprisingly unexpected because peer-assessors attached the most considerable importance to Content; by contrast, Grammar was the least attended criterion for peer-assessors. This finding does not fit the Iranian context and goes against some of the studies conducted in the Iranian setting as outlined in the preceding paragraph. Possible explanations for this tendency of peer-assessors in this study may be attributed to the following factors. First, raters' ability to understand and respond to the characteristics of a rating criterion may affect their beliefs about the criteria. In other words, raters do not have a clear understanding of a certain criterion (Marefat & Heydari, 2016, p. 32). Second, raters may find one criterion difficult and then attach less importance to it, or perceive one criterion easy and pay more attention to it (Lee, 2009). Peer-assessors in the present study may have perceived content easy and grammar difficult, thereby attaching the highest importance to the former and the lowest importance to the latter.

5. Conclusion and Implications

As for severity in peer-assessors when assessing the essays of their peers, the results showed that advanced peer-assessors had more variability in their severity compared to intermediate peer-assessors. Second, the majority of peer-assessors were, on average, more severe than lenient. Third, the

results of an independent *t*-test revealed that there was no statistically significant difference between the ratings of intermediate and advanced peer-assessors in rating, implying that the level of proficiency was not related to the rating of the essays. The final finding was that *task achievement* was the most attended assessment criterion, but *grammatical range and accuracy* was the least attended assessment criterion.

According to the findings of the present study, it may be safe to conclude that the proficiency of peer-assessors may not be an important factor to invest in because severity measures of both intermediate and advanced peer-assessors, on average, were not statistically significant. This may imply the ratings for these two groups of assessors can be used for achievement purposes when language learners are tasked with evaluating the products of their peers. Further, as the findings of the study showed, regardless of proficiency level of peer-assessors, most peer-assessors were severe.

Severity of peer-assessors may stem from some other factors. Personality traits, rater training, gender, and rating strategies may affect severity measures. For example, Fahim and Bijani (2011) noted that "training reduced raters' severity and harshness to a great extent but did not eliminate it" (p. 11). One tentative conclusion which can be drawn may have to do with longer duration of training programmes to reduce significantly severity differences.

Peer-assessors were more concerned about content and organisation and less attentive to grammar and vocabulary. Although both surface features (e.g., grammar) and discourse features (e.g., content and organisation) can be used for rating purposes, discourse features should be prioritised when peer-assessors are engaged in assessing the works of their peers. The tendency of putting more importance on content and organisation was that raters had "more emphasis on how well a writer presents what he/she wants to convey" (Lee, 2009, p. 393).

The findings of the study can be beneficial for training purposes to instruct peer-assessors to best use the rating scale criteria and the guidelines about how to rate essays. Actually, teachers can add new trends to traditional testing and exams. This study can provide support for teacher and students' more cooperative and communicative work in classrooms. Educational setting such as language institutes, schools, and universities can also take advantage of the findings of present study. They can use peer assessment to take responsibility for taking part in assessment of their classmates and to change a traditional way of evaluation (teacher-to-students) to peer-assessment evaluation.

References

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied measurement in Education, 4*(4), 275-288.

Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing, 8*(3), 225–241.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5*(1), 7-74.

Boud, D. (1989). The role of self-assessment in student grading. *Assessment in Higher Education, 14*(1), 20-30.

Boud, D. (1995). Assessment and learning: Contradictory or complementary? In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 35–48). Kogan Page Limited.

Boud, D. (2003). *Enhanced learning through self-assessment*. Kogan Page.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413-426.

Boud, D., Cohen, R., & Sampson, J. (Eds.) (2001*). Peer learning in higher education: learning from and with each other.* Kogan Page.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. Routledge.

Brown, H. (2004). Language assessment: *Principles and classroom practices*. Longman

Brown, J. B., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653- 675.

Brown, S., & Glasner, A. (Eds.) (1999). *Assessment matters in higher education: choosing and using diverse approaches*. SRHE and Open University.

Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. Kogan Page.

Brown, S., Race, P., & Rust, C. (1995). Using and experiencing assessment. In P. Knight (Ed.) *Assessment for learning in higher education (*pp.75–85). Kogan Page.

Cheng, W., & Warren, M. (1997). Having second thoughts: student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233-239.

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

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

Connor-Linton, J. (1995). Cross-cultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*, 99–115.

Dancer, W. T., & Dancer, J. (1992). Peer rating in higher education. *Journal of Education for Business, 67*(5), 306–310.

DiGiovanni, E., & Nagaswami, G. (2001). Online peer review: An alternative to face-to-face? *ELT journal, 55*(3), 263-272.

Dochy, F. J., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in educational evaluation, 23*(4), 279-298.

Dochy, F., Segers, M. & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331–350.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2007). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196.

Esfandiari, R. (2015). Rater Errors among Peer-Assessors: Applying the Many-Facet Rasch Measurement Model. *Iranian Journal of Applied Linguistics*, *18*(2), 77-107.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111-131.

Fahim, M, & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1), 2251-7324.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self-assessments. *Assessment and Evaluation in Higher Education, 11*(2), 146-166.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and Training International, 32*(2), 175–187.

Falchikov, N. (2003). Involving students in assessment. *Psychology Learning and Teaching, 3(2),* 102–108.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment: A meta-analysis comparing peer and teacher remarks. *Review of Educational Research, 70,* 287-322.

Gibbs, G. (1999). Using assessment strategically to change the way students learn, In S. Brown & A. Glasner (Eds.), *Assessment matters in higher Education: Choosing and using diverse approaches (*pp. 41–53*).* SRHE & Open University Press.

Goldfinch, J. M., & Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment & Evaluation in Higher Education, 15*(3), 210–225.

Hargreaves, L., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal, 39*(1), 69–96.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Association for Supervision and Curriculum Development.

Huerta-Macías, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*(1), 8–11.

Kolomuç, A. (2017). Subject-specific science teachers' views of alternative assessment. *Asia-Pacific Forum on Science Learning and Teaching*, *18*(1), 124-135.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348.

Leach, L., Neutze, G. & Zepke, N. (2001). Assessment and empowerment: some critical question. *Assessment and Evaluation in Higher Education, 26*(4), 293–305.

Lee, H. K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific Education Review, 10*(3), 387–397.

Lee, I. (2004). Error correction in L2 secondary writing classroom: The case of Hong Kong. Journal of Second Language Writing, 13(4), 285–312.

Li, H., Bialo, J. A., Xiong, Y., Hunter, C.V., Guo, X. (2021). The effect of peer assessment on non-cognitive outcomes: A meta-analysis. *Applied Measurement in Education*, *34*(3), 179-203.

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2011*). FACETS* (Version 3.68.1) [Computer Software]. Chicago, IL: MESA Press.

Loddington, S. (2008). Peer assessment of group work: A review of the literature. Retrieved from http://webpaproject.lboro.ac.uk/files/WebPA_Literature%20review%20.pdf

Long, M. (1985). Input and second language acquisition theory. In S. M. Gass & C. G. Maddan (Eds.), *Input in second language acquisition* (pp. 337-393). Newbury House.

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing, 27*(1), 24-36.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100.

McDowell, L. & Sambell, K. (1999). The experience of innovative assessment: students' perspectives. In S. Brown & A. Glasner (Eds.),

Assessment matters in higher education: choosing and using diverse approaches (pp. 71–82). SRHE & Open University Press.

Mendonca, C. O., & Johnson, K. E. (1994). Peer review negotiations: Revision activities in ESL writing instruction. *TESOL Quarterly, 28*(4), 745-769.

Minzi, L., & Zhang, X. (2021). A meta-analysis of self-Assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189-218.

Myford, C. M., &Wolfe, E. W. (2004). Detecting and measuring rater effects using many facet Rasch measurement: Part II. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460–517). JAM Press.

Nakamura, Y. (2002). *Teacher assessment and peer assessment in practice (Educational Studies 44).* International Christian University. (ERIC Document Reproduction Service No. ED464483).

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21*(3), 239–250.

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 25*(1), 23-38.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309-323.

Phakiti, A. (2003). A closer look at gender and strategy use in L2 reading. *Language Learning, 53*(4), 649–702.

Ross, S. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics, 26*(3), 317–342.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209.

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

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54.

Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. *RELC Journal, 40*(2), 149–171.

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: A case study in the school of music, Kingston University. *Assessment & Evaluation in Higher Education, 22*(4), 371–383.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environment Research, 1,* 293-319.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2005). Trends in peer learning. *Educational psychology, 25*(6), 631-645.

Topping, K. J. (2009). Peer assessment. *Theory into Practice, 48*(1), 20–27.

Topping, K. J. (2010). Peers as a source of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 69–75). Routledge.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education, 25*(2), 146–169

Vygotsky, L. S. (1987). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), p. 370.

Zhan, Y. (2021). What matters in design? Cultivating undergraduates' critical thinking through online peer assessment in a Confucian heritage context. *Assessment & Evaluation in Higher Education, 46*(4), 615-630.

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.