Investigating the Effect of Self-, Peer-, and Teacher Assessment in Second Language Writing over Time: A Multifaceted Rasch Approach

Sadollah Ravand
M.A.,Vali-e-Asr University of Rafsanjan
s4ravand@gmail.com

Hamdollah Ravand
Assistant Professor,
Vali-e-Asr University of Rafsanjan
ravand@vru.ac.ir

Akbar Abbasi
Assistant Professor,
Vali-e-Asr University of Rafsanjan
abbasi@vru.ac.ir

Abstract

This study investigated the accuracy of scores assigned by self-, peer-, and teacher assessors over time. Thirty-three English majors who were taking the paragraph development course at Vali-e-Asr University of Rafsanjan (VRU) and two instructors who had been teaching essay writing for at least two years at VRU, participated in the study. After receiving instructions on paragraph development, participants were trained for a session on how to rate the paragraphs. For three sessions the students were given topics to write about and were asked to rate their own and one of their peers' papers for mechanics, grammar and choice of words, content development, and organization. The teachers also rated the paragraphs according to the same criteria. Multifaceted Rasch measurement was employed to analyze the data. The results showed different patterns of performance for the subjects rated by different raters at the beginning of the experiment. However, rater bias showed significant decrease across time. The results of the study have useful implications for language teachers especially in portfolio assessment where self and peer assessment provide invaluable help.

Keywords: EFL Writing, Multifaceted Rasch Measurement, Peer-Assessment, Self-Assessment, Teacher Assessment

Received: February 2015; Accepted: November 2015

1. Introduction

It is generally agreed that to assess student's learning, proficiency, and knowledge, teachers need to use a range of assessments (Orsmond, Merry, & Pope, 2005; Reiling, 2000). Nonetheless, in traditional classroom settings, the teacher is the sole evaluator. The traditional approach is appropriate, when students take an objective test; however, the use of a single assessor in performance tests, such as essays, oral presentations, and role-plays can lead to biased evaluations (Matsuno, 2009). In order to compensate for the limitations of teacher assessment, alternative means of assessment have been developed in the field of education.

Among the alternative means of assessment, self- and peer-assessment have intrigued much attention due to a growing emphasis on learner independence and autonomy (Sambell, McDowell, & Sambell, 2006). According to Hargreaves, Earl, and Schmidt (2002)

> These alternative means of assessment motivate students to take more responsibility for their own learning, to make assessment an integral part of the learning experience, and to embed it in authentic activities that recognize and stimulate students' abilities to create and apply a wide range of knowledge, rather than simply engaging in acts of memorization and basic skill development. (p. 70)

Moreover, self- and peer-assessment have been considered as having critical pedagogical values. According to Brown and Hudson (2002), less time is required to conduct self-assessment in the classroom. In addition, involvement of students in the process of assessment, can result in learner autonomy and higher motivation (Alibakhshi, 2013; Dickinson, 1987; Harris, 1997; Oscarson, 1989).

In spite of the aforementioned benefits of self- and peer-assessment, they have not been widely used in educational settings. This is probably due to the fact that the reliability of self- and peer-assessment, and the ability of the learners to accurately assess themselves and their peers is doubted. This uncertainty of teachers about the learners' ability to do self-, and peer-assessment reliably, has been further complicated by contradictory results of the studies on the reliability of self- and peer-assessment (Oscarson, 1989; Patri, 2002). Some researchers have found self-assessments to be more reliable than peer-assessments (Falchikov, 1986; Jiliang & Kun, 2007; Longhurst & Norton, 1997). By contrast, Saito and Fujita (2004), Matsuno (2009), Esfandiari and Myford (2013) found that self-assessment does not have sufficient reliability and concluded that it is of limited utility as a part of formal assessment. Instead, they recommended using peer-assessment in writing classes for formative assessment.

Although some studies have investigated the reliability of self- and peer-assessment, relatively few studies have investigated the effectiveness of these assessment types in English as a foreign language (EFL) settings (e.g., Cheng & Warren, 1997; De Grez, Valcke, & Roozen, 2012; Esfandiari & Myford, 2013; Matsuno 2009; Nakamura, 2002; Patri, 2002; Saito & Fujita, 2004).

Additionally, empirical studies on the effect of self-, and peer-assessment over time are very scanty (e.g., Butler & Lee, 2010). Moreover, few researchers have compared severity differences among self-, peer-, and teacher assessors (e.g., Esfandiari & Myford, 2013; Matsuno, 2009; Saito & Fujita, 2004). Most of the researchers in this field have mainly used simple correlation and traditional true-score approach. Few studies have used Multifaceted Rasch Measurement (MFRM) to compare ratings assigned by different rater types (e.g., Esfandiari & Myford, 2013; Matsuno, 2009; Nakamura, 2002; Saito & Fujita, 2004). In the

present study, MFRM was used to compare self-, and peer-assessment with teacher assessment. In contrast to the traditional approaches, MFRM can show essay quality, rater severity, item difficulty, and the degree to which raters are internally and externally consistent.

In the subsequent section, some major related studies in the literature are reviewed.

2. Review of the Literature

Different researchers have defined self-assessment differently. For example, Andrade, Du, and Mycek (2010), defined self-assessment as "a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise accordingly" (p. 3). Peer-assessment can be defined as "an arrangement for peers to consider the level, value, worth, quality or successfulness of the products or outcomes of learning of others of similar status".(Topping, Smith, Swanson, & Elliot, 2000, p. 150).

Some researchers (e.g., Blue, 1994; Cheng & Warren, 2005; Falchikov & Goldfinch, 2000; Oscarson, 1989; Saito & Fujita, 2009; Topping, 2009) have enumerated indispensable values for self-, and peer-assessment. Saito and Fujita (2004), for example, argued that self-, and peer-assessment promote responsibility for managing the assessment process, thus increasing responsibility for learning. London and Tornow (1998) maintained that feedback from different sources increases self-awareness, and noticing the gap between self- and other students' understanding, encourages the students to fill the gap, which in turn leads to further learning. Brown (1998) held that assessing peers makes the students more sensitive to the evaluation criteria, and promotes self-reflection as a result.

Most of the existing studies that have been conducted on self-, and peer-assessment, have focused on the reliability and educational benefits of these assessment instruments; however, mixed results have been reported. It is still doubted that whether self-, and peer-assessment could be used to make important decisions in educational settings. In the following, some studies relevant to the concerns of the present study will be reviewed.

In a study involving university students in Iran, Esfandiari and Myford (2013) compared severity of self-, peer-, and teacher- assessments in foreign language writing. They found that on average teacher assessors rated more severely while self-assessors rated more leniently. Peer-assessors turned to be halfway between those two assessor groups. They attributed these differences to the influence of cultural mores on students' abilities to self-assess and peer assess. In a similar study, but in a different culture, Matsuno (2009) compared self-, and peer-assessments with teacher assessments in university writing classes. He found that in comparison with self-, and peer-assessors, teacher assessors were neither lenient nor severe. Peer-assessors produced fewer biased interactions compared with the self-, and teacher assessors. Based on the results of the study, he recommended that, in some contexts, peer-assessment can be used in writing classes. He concluded that "self-assessment was somewhat idiosyncratic and therefore of limited utility as a part of formal assessment" (p.75). Matsuno's finding mirrors the finding of the study by Esfandiari and Myford (2013) in that in both studies self-assessors were found to be significantly more lenient than peer-, and teacher assessors. In sharp contrast, some other researchers (e.g., Falchikov, 1986; Lunghurts & Norton, 1997; and Jiliang & Kun, 2007) found that self-assessments were more reliable than peer assessments.

Examining the characteristics of peer ratings of Japanese students, Saito and Fujita (2004) studied self-, peer-, and teacher assessments. The results from their analysis showed that self-raters were the most lenient but also the most severe raters compared to peer-, and teacher raters. Compared to the other two groups, peer-raters were lenient on average. Teacher raters were found to be severer than peer-raters, but they were less severe than self-raters. They also found a strong positive correlation between peer-ratings and teacher ratings. Comparing the ratings of peer-raters and teacher raters, Saito (2008) found similar results to those of Saito and Fujita (2004). He found that teacher raters were severer than the peer-raters in rating all aspects of the oral presentations except the visual aspects. In a later study, Saito and Fujita (2009) compared teachers' and peer-assessors' rating of students' group presentation and they found a high positive correlation between their ratings. The finding that peer-assessors and teacher assessors have similarity in scoring and they have no correlation with self-assessors is consistent with studies of second language oral presentations (Patri, 2002; Yamashiro, 1999).

Lindblom-Ylänne, Pihlajamaki, and Kotkas (2006) compared the results of self-, peer-, and teacher assessment of student essays. They investigated the students' experience of self-, and peer-assessment processes, as well. The results showed that self-, peer-, and teacher assessments were quite similar to each other. In general, both the teachers and the students had positive experiences of self-, and peer-assessment. The results of self-assessment in Lindblom-Ylänne et al.'s study were very similar to the results of peer-, and teacher assessment in the literature extensively reviewed by Dochy, Segers, and Sluijsmans(1999). Furthermore, their results are in contrast with those of Esfandiari and Myford (2013), and Falchikov and Boud (1989), who found that self-assessment grades tended to be higher than peer-, and teacher assessment

grades. Moreover, the results did not support the tendency of over-rating in peer-assessment that was reported in previous studies (e.g., MacKenzie, 2000; Magin & Helmore, 2001; Topping et al., 2000).

In a study on self-assessment, LeBlanc and Painchaud(1985) found that students' self-assessment could be reliable enough to be used for placement purposes. Similar corroborating results of the reliability and validity of self-, and peer-assessment have also been found by Ross (1998), Cheng and Warren (2005), Patri (2002), and Saito and Fujita (2004). In another study, Butler and Lee (2010) examined the effectiveness of self-assessment of English performance among young learners of English. They found that the students improved their ability to self-assess their performance over time.

3. Research Questions

The present study aims to investigate the following questions:

1. How do self-, and peer-assessment differ compared to teacher assessment in the level of severity/strictness?
2. Do raters rate differently over time?

4. Method

4.1. Participants

The participants of this study consisted of 33 female students from two intact classes, who functioned as self-assessors, and peer-assessors. The student assessors were senior English majors at Vali-e-Asr University of Rafsanjan. Their ages ranged from 18 to 25. They all were native Farsi-speakers. Most of them had studied English language in language institutes before entering the university. None of them, however, had had any extensive writing instruction

experience before. In addition to the student participants, the instructor of the course and a teacher who had had the experience of teaching writing classes for at least two years, acted as the teacher assessors.

4.2. Instrument

The rating scale employed to score the students' paragraphs was the ESL composition profile validated by Aryadoust (2010) via a structural equation modeling (See Appendix A). This instrument contained criteria that focused on the following key concepts: (1) Arrangement of Ideas and Examples (AIE) and Communicative Quality (CQ). AIE concerns the appropriate tone of the text and genre, exemplification, arrangement of ideas, completeness of responses to the prompt, and topic relevance. The Communicative Quality (CQ) or Coherence and Cohesion (CC) includes elements of argument in which components of causality and coherent presentation of ideas are crucial. (2) Sentence Structure Vocabulary (SSV), which encompasses the employment of appropriate vocabulary, and correct spelling, punctuation and syntax. Each item contained a five-point rating scale which ranged from (1) poor, (2) fair, (3) good, (4) very good and (5) excellent.

4.3. Procedure

The study was conducted in two paragraph writing classes, instructed by the same instructor. Between the first and sixth weeks of the semester, the students received instructions concerning paragraph development such as mechanics, grammar, and choice of words, content development, and organization. Following the seventh session, students were given a topic from their books to write paragraphs on. Prior to rating the first assignment, the students had a

two-hour training session in which they were instructed regarding how to assess the paragraphs according to the above mentioned criteria. First, the instructor elaborated on the items of the rating scale and the marking criteria in detail and gave the students all the necessary guidelines about how to rate a paragraph. Afterwards, the instructor displayed a sample paragraph produced by one of the students on the screen and rated it according to the guidelines and rating criteria. Then, students received another sample paragraph and rated it on their own according to the guidelines. During this phase of the training, the instructor went over the class and monitored their ratings and explained any unclear points. The first writing assignment was due the seventh session. Following the training session, each student rated his/her own paragraph. The instructor advised them to assess as accurately as possible. When they had completed this task, the students were asked to peer-assess one of her classmates' anonymous papers. The students rated their peers' paragraphs according to the same procedures. After they finished rating the paragraphs, the instructor collected the rating sheets. The students repeated the same self-, and peer-assessment on the eleventh and fifteenth sessions, and the data needed for this study were collected from these three rating sessions.

The second teacher assessor also had the same assessment training, but in a separate session. After assessing the paragraphs by the students, the teacher assessors used the same procedures to assess the students' paragraphs. It should be noted that the student assessors and the teacher assessors rated the paragraphs independently.

5. Results

Question 1: How do self-, and peer-assessment differ compared to teacher assessment in the level of severity?

As presented in Figure 1, the facets map displays visually the relative abilities of examinees, the relative severity of the raters, and the relative difficulty of the tasks. The map displays all facets of the analysis, summarizing key information about each facet. The first column displays the logit scale, which ranges from 1 to -2 logits. The average task difficulty has been set at 0 logit, so that tasks with negative signs are easier than average, and those with positive signs are more difficult than average. With persons, the higher on the scale, the more able; and with raters, the higher on the scale, the severer. The second column shows the three occasions that the subjects wrote a paragraph. In the third column the severity of each of the five raters is given.

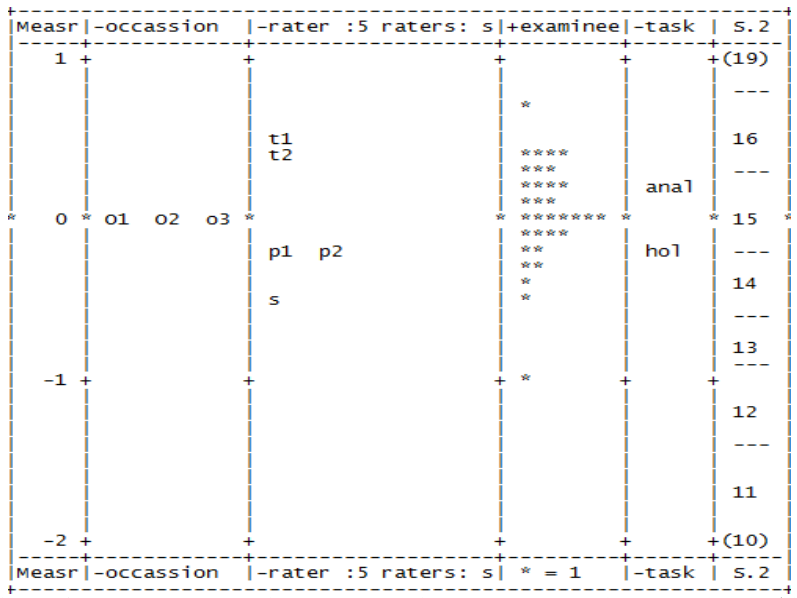

Figure 1.*Facets Map of Different Occasions, Rater Severity, Examinee Ability*

As indicated on the map, the raters indeed differ in their level of severity, by as much as less than 2 units on the logit scale. Teacher assessors tended to rate rather more severely on average, while self-assessors tended to rate rather more leniently on average. Peer-assessors appeared midway between those two

assessor types. Teacher assessors were the severest assessors, but self-assessors were the most lenient ones. The fourth column displays examinees' abilities. In general, the students' abilities correspond to the raters' severity.

Figure 2 shows the student writers' abilities in ascending order. Negative values show low ability and positive values show high ability. For example, as shown in the last column, student 31 with the measure value of -1.02 has the lowest level of ability and student 22 with the measure value of .70 has the highest level of ability. As indicated by the *t*-values (zstd) in columns 8 and 10, all student writers fit the model. None of the *t*-values exceeds the accepted boundaries of ±2.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu examinee
38	3	12.67	12.33	-1.02	.39	.98	.2	.95	.1	1.16	.92	.80	31 31
143	10	14.30	13.72	-.49	.20	.64	-.8	.64	-.8	1.54	.91	.57	32 32
164	11	14.91	14.23	-.43	.20	1.72	1.5	1.79	1.7	-.02	.35	.47	23 23
118	8	14.75	14.23	-.30	.23	.58	-.8	.56	-.9	1.47	.88	.55	30 30
132	9	14.67	14.28	-.28	.22	1.11	.3	1.04	.2	.98	.30	.55	33 33
195	13	15.00	14.44	-.21	.19	.64	-.9	.63	-.9	1.35	.68	.55	19 19
163	11	14.82	14.46	-.21	.20	1.39	.9	1.34	.8	.70	.74	.53	14 14
137	9	15.22	14.61	-.14	.23	.96	.0	.90	.0	1.15	.70	.55	18 18
132	9	14.67	14.62	-.14	.22	.90	.0	.97	.0	.84	.67	.52	2 2
169	11	15.36	14.79	-.06	.21	.69	-.6	.71	-.6	1.34	.61	.56	10 10
198	13	15.23	14.81	-.06	.19	.63	-.9	.61	-1.0	1.40	.07	.55	8 8
138	9	15.33	14.90	-.02	.23	1.01	.1	1.03	.2	1.01	.61	.56	20 20
151	10	15.10	14.91	-.01	.21	.63	-.8	.65	-.7	1.35	.24	.55	4 4
185	12	15.42	14.94	.00	.20	1.12	.4	1.15	.4	.86	.48	.55	9 9
169	11	15.36	14.98	.02	.21	1.22	.6	1.15	.4	.82	.64	.54	16 16
201	13	15.46	14.99	.02	.19	1.03	.2	.95	.0	1.08	.62	.52	25 25
155	10	15.50	15.02	.04	.22	1.36	.8	1.39	.9	.58	.05	.54	17 17
155	10	15.50	15.02	.04	.22	.63	-.8	.61	-.8	1.37	.48	.54	24 24
156	10	15.60	15.15	.10	.22	.58	-.9	.61	-.8	1.43	.59	.54	27 27
186	12	15.50	15.19	.11	.20	.88	-.1	.91	.0	1.10	.44	.52	28 28
172	11	15.64	15.22	.13	.21	1.85	1.7	1.86	1.7	-.07	.27	.55	7 7
188	12	15.67	15.33	.17	.20	1.34	.8	1.44	1.0	.58	.34	.54	6 6
79	5	15.80	15.35	.19	.32	1.27	.5	1.23	.5	.79	.97	.58	26 26
238	15	15.87	15.59	.21	.19	.76	-.5	.81	-.4	1.23	.48	.42	11 11
222	14	15.86	15.45	.23	.19	.77	-.5	.80	-.4	1.19	-.14	.52	15 15
192	12	16.00	15.57	.29	.21	.89	-.1	.86	-.2	1.11	.24	.53	1 1
191	12	15.92	15.64	.32	.21	.99	.1	.96	.0	.89	.51	.51	29 29
175	11	15.91	15.64	.32	.22	.45	-1.4	.41	-1.6	1.55	.56	.54	3 3
192	12	16.00	15.73	.37	.21	.75	-.5	.76	-.5	1.26	.30	.51	5 5
161	10	16.10	15.75	.38	.23	1.32	.7	1.44	1.0	.68	.38	.50	21 21
226	14	16.14	15.76	.38	.20	1.18	.5	1.19	.5	.88	.38	.51	12 12
209	13	16.08	15.76	.39	.20	.99	.1	1.03	.2	.93	.49	.50	13 13
168	10	16.80	16.35	.70	.26	1.59	1.2	1.54	1.1	.32	.51	.47	22 22
166.6	10.8	15.40	14.99	.03	.22	1.00	.0	1.00	.0		.49		Mean (Count: 33)
39.5	2.4	.71	.73	.32	.04	.34	.8	.35	.8		.25		S.D. (Population)
40.1	2.4	.72	.74	.32	.04	.35	.8	.36	.8		.26		S.D. (Sample)

Figure 2. *Examinees Measurement Report*

Values of *t* outside the range of approximately -2 to +2 are said to indicate significant departure from the expectations of the model. Values larger than +2 indicate significant underfit; values below -2 indicate significant overfit.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N rater :5 raters: s
1174	71	16.54	16.09	-.51	.09	.73	-1.7	.73	-1.7	1.27	.17	.34	18.3	16.0	1 s
850	53	16.04	15.52	-.21	.10	1.33	1.5	1.36	1.7	.58	.28	.39	13.8	17.8	3 p2
1202	75	16.03	15.45	-.20	.08	1.16	1.0	1.17	1.0	.82	.23	.34	17.1	17.7	2 p1
1013	69	14.68	14.00	.42	.08	.91	-.4	.92	-.4	1.10	.46	.40	14.2	16.0	5 t2
1259	87	14.47	13.79	.50	.07	.89	-.7	.88	-.7	1.12	.53	.40	14.3	15.0	4 t1
1099.6	71.0	15.55	14.97	.00	.09	1.01	-.1	1.01	.0		.33				Mean (Count: 5)
149.2	11.0	.82	.91	.39	.01	.21	1.2	.22	1.3		.14				S.D. (Population)
166.8	12.2	.92	1.02	.44	.01	.24	1.3	.25	1.4		.15				S.D. (Sample)

Figure 3. ***Raters Measurement Report***

Figure 3 shows the raters in ascending order of severity. Self-raters were the most lenient, but teacher raters were the severest. Based on a useful rule of thumb (McNamara, 1996) values in the range of approximately 0.7 to 1.3 are acceptable and satisfy the expectations of the model. Thus, the performance of all raters fit the model. Values greater than 1.3 show significant underfit, that islack of predictability; values below 0.7 show significant overfit, that is too much predictability. For example, the infit measure (infit MnSq) for peer-assessor 2 is 1.33 which is slightly underfit, but as the *t*-value 1.5 did not exceed the accepted boundaries of±2, no misfit was observed.

Question 2: Do raters rate differently over time?

Figure 4 displays rater-bias over occasion. Based on the overall severity of the raters, the model expected that the overall score that peer 2 on Occasion 2 assigned to all the writers (expected score) should be 225, but what we observed (observed score) was 216. It indicated that she rated more severely than expected. There was a discrepancy between the expectation of the model and the observed score. This discrepancy expressed in logit, is -.37(bias size). We wanted to see whether this difference was significant between what the model expects and what we observed. The *t*-value (-1.98) was below 2, so it was not

significant. For peer 1(p1) on Occasion 2(O2) the expected score was 448 but the observed score was 433. It indicated that they rated more severely. This discrepancy was significant because the *t*-value (-2.18) exceeded the accepted boundaries of ±2. The *t*-value for self-rater on Occasion 1 was 2.30 which was significant. Overall, the majority of the raters did not show bias over occasion.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	occassion Sq	N	oc	rater measr	N	ra	:5 ra measr
216	225.96	14	-.71	-.37	.19	-1.98	13	.0694	1.0	1.1	8	2	O2	.00	3	p2	-.21
433	448.79	28	-.56	-.29	.13	-2.18	27	.0382	1.0	1.0	5	2	O2	.00	2	p1	-.20
305	314.04	19	-.48	-.27	.17	-1.61	18	.1258	.6	.6	2	2	O2	.00	1	s	-.51
359	366.36	22	-.33	-.21	.16	-1.26	21	.2199	.7	.7	3	3	O3	.00	1	s	-.51
405	419.54	29	-.50	-.20	.12	-1.73	28	.0946	.7	.7	11	2	O2	.00	4	t1	.50
310	311.07	21	-.05	-.02	.14	-.15	20	.8794	1.1	1.1	15	3	O3	.00	5	t2	.42
335	334.88	23	.01	.00	.14	.02	22	.9875	.8	.8	13	1	O1	.00	5	t2	.42
368	367.10	25	.04	.02	.13	.12	24	.9071	.9	.9	14	2	O2	.00	5	t2	.42
366	364.51	25	.06	.03	.13	.19	24	.8490	.8	.9	12	3	O3	.00	4	t1	.50
325	323.14	20	.09	.05	.17	.31	19	.7614	.9	.9	6	3	O3	.00	2	p1	-.20
421	415.42	26	.21	.12	.15	.80	25	.4339	1.7	1.8	7	1	O1	.00	3	p2	-.21
488	475.02	33	.39	.17	.11	1.46	32	.1550	.9	.9	10	1	O1	.00	4	t1	.50
213	208.61	13	.34	.19	.21	.89	12	.3893	.5	.5	9	3	O3	.00	3	p2	-.21
444	430.06	27	.52	.29	.15	1.94	26	.0636	1.2	1.2	4	1	O1	.00	2	p1	-.20
510	493.58	30	.55	.35	.15	2.30	29	.0288	.5	.5	1	1	O1	.00	1	s	-.51
366.5	366.54	23.7	-.03	-.01	.15	-.06			.9	.9	Mean (Count: 15)						
84.1	80.55	5.5	.39	.21	.03	1.38			.3	.3	S.D. (Population)						
87.1	83.38	5.7	.41	.22	.03	1.43			.3	.3	S.D. (Sample)						

Figure 4. ***Rater Bias***

5.1. Bias Analysis

It is possible that the raters may display particular patterns of harshness or leniency in relation to only one group of examinees, not others, or in relation to particular tasks or occasions, not others. Multifaceted analysis compares expected and observed values in a set of data. Fit statistics (for raters, persons, tasks and other facets) summarizes the extent to which the differences between expected and observed values are within a normal range. (Values outside the range of approximately +2 to -2 suggest significant bias).

We intended to see whether there was a pattern in the bias from Occasion 1 to Occasion 3. The expectation was that the bias reduces from Occasion 1 to Occasion 3. This pattern was evident in most of the raters (except for peer 2 that negligibly increased).

Table 1. *Rater Bias over Occasion*

Rater	Occasion	Bias	Rater	Occasion	Bias	Rater	Occasion	Bias	Rater	Occasion	Bias	Rater	Occasion	Bias
T1	1	.17	**T2**	1	.00	**P1**	1	.29	**P2**	1	.12	**S**	1	.35
T1	2	-.20	**T2**	2	.02	**P1**	2	-.29	**P2**	2	-.37	**S**	2	-.27
T1	3	.03	**T2**	3	-.02	**P1**	3	.05	**P2**	3	.19	**S**	3	-.21

Table 1 shows rater bias over occasion. These results suggested that the majority of the raters showed less bias from occasion 1 to Occasion 3. For example, as shown in Figure 4, peer 2 on Occasion 2 rated more severely, but this bias was not statistically significant as indicated by the *t*-value and probability which was not below .05 and the *t*-value did not exceed the expected boundaries of ±2. Peer 1 on Occasion 2 rated more severely and the bias was significant as indicated by the *t*-value which exceeded the boundaries of ±2 and the probability level was below .05. Self-assessors on Occasion 1 rated more leniently and this bias was significant as indicated by the *t*-value which exceeded the accepted range of ±2 and the probability level which was below .05. Overall, most of the raters did not show bias over occasion.

The same information regarding rater bias over occasion is displayed in figure 5. Except teacher 2, all the other raters showed the same bias pattern. They were more lenient on Occasion 1, severer on Occasion 2 and again more lenient on Occasion 3.

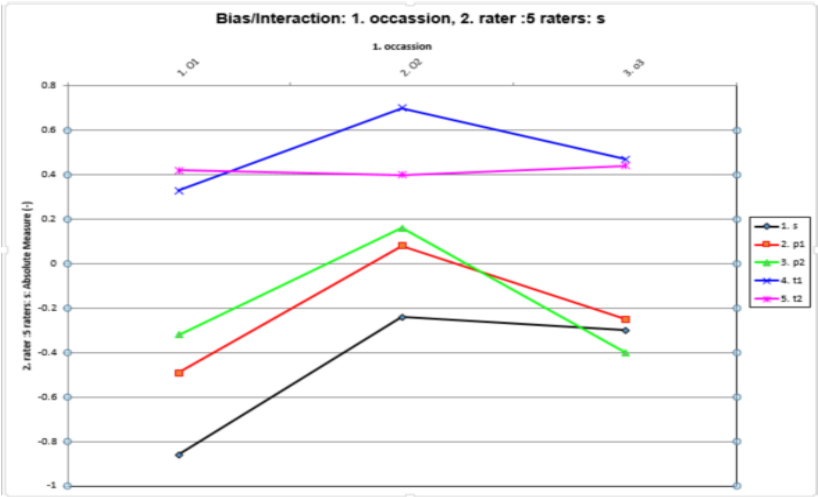

Figure 5. *Rater Bias over Occasion*

6. Discussion

In the present study we investigated the performance of self-, and peer-assessors in comparison with teacher assessors when rating paragraphs. MFRM was used to measure and detect severity differences among the three assessor types.

Regarding the first research question, the results of the data analysis indicated that, of the three assessor types, teacher assessors were the severe starters, while self-assessors tended to rate themselves more leniently on average. Peer-assessors appeared midway between those two assessor groups but their ratings tended to be more similar to those of the self-assessors than to those of the teacher assessors.

The findings of this study mirrored some of those from the previous studies. For example, comparing self-, peer-, and teacher assessment, our results support those of Esfandiari and Myford (2013) and Sullivan and Hall (1997),

who like us, found that self-assessors tended to rate themselves more leniently. However, our results are in sharp contrast with those of Leach (2000), Brown (2005), Chen (2008), and Matsuno (2009), who found that the self-assessors tended to rate themselves more severely. The tendency of self-assessors to underrate themselves can be attributed to the cultural value of modesty among the test takers. But in an Iranian context, as Esfandiari and Myford (2013) argued, "students do not share that cultural value of modesty. Student evaluations tend to be norm-referenced" (p.125). Therefore, self-assessors are inclined to assign ratings that are higher than those that they would assign to their peers' writings.

Comparing peer-, and teacher assessors, we found that peer-assessors tended to be closer to the ratings of the teacher assessors, but more lenient than teacher assessors. This finding supports those of Saito and Fujita (2004) and Saito (2008) who also found peer-assessors were more lenient than teacher assessors. In sharp contrast, Nakamura (2002) found that teacher assessors were more lenient than peer-assessors. As Esfandiari and Myford (2013) argued the leniency of the peer-assessors towards their peers could be attributed to the cultural beliefs. Students do not want to be critical of their classmates and think if they assign low ratings to their classmates, this may ruin friendship and cause animosity. Another possible reason could be attributed to Islamic teachings, which encourages one to first care about their neighbors.

In our study, we also investigated the performance of self-, peer-, and teacher assessors over time. As indicated by the results of the bias analysis, the majority of the raters showed the same bias pattern. Overall, they rated the writings more leniently on Occasion 1, more harshly on Occasion 2, and again more leniently on Occasion 3. They showed a decreasing bias pattern from occasion 1 to occasion 3. Generally, most of the raters did not show bias over

occasion. A possible explanation for this finding could be related to the skill that the raters developed over time. As they rated more paragraphs, their rating skills improved and they showed less bias over time.

7. Conclusion

Using a Multifaceted Rasch Measurement, this study has reported on a quantitative investigation of self-, peer-, and teacher assessments of paragraphs written by Iranian students studying English language at undergraduate level. Some important findings emerged. First, of the three assessor types, self-assessors were the most lenient, and teacher assessors were appeared the severest raters. Second, when scoring analytically, raters were severer, and they were more lenient with holistic scoring. Third, most of the raters did not show bias over occasion, self and peer bias pattern from Occasion 1 to Occasion 3 had a decreasing trend. It indicated that more rater training sessions and more writing samples over more occasions can result in little bias in self-, and peer-assessment.

The findings of this study have significant pedagogical implications for EFL writing teachers. Language teachers are not recommended to use self-, and peer-assessments in rating formal and high-stakes assessments and for summative purposes. These alternative means of assessment can be used for portfolio assessment and also in low-stakes, formative decision making contexts.

The generalizability of the results of the present study is limited in that we used a unisex sample of a relatively small size and the subjects were not selected randomly. In addition, we had just a single rater training session. The raters were trained only for two hours and they rated only one sample paragraph for practice. Perhaps more training sessions and more practice in

rating paragraphs might have helped them to internalize the guidelines and rating criteria more deeply and have been more adept at assessing the paragraphs more objectively, and as a result, have produced little bias in self-, and peer-assessment. Moreover, we elicited students' writing performance on just three occasions and we did not control for the level of proficiency of the subjects.

References

Alibakhshi, G. (2013). Construction and Validation of Self-assessment inventory: A case of Iranian tertiary students. *Research in Applied Linguistics Studies, 4*(2), 93-109.

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17*(2), 199-214. doi: 10.1080/09695941003696172

Aryadoust, V. (2010). Investigating writing sub-skill in testing English as a foreign language: A structural equation modeling study. *TESL-EJ, 13*(4).

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman London.

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment & Evaluation in Higher Education, 33*(4), 431-444. doi: 10.1080/02602930701562775

Blue, G. M. (1994). Self-Assessment of foreign language skills: Does it work? *CLE Working Papers, 3*, 18-35.

Brown, A. (2005). Self-assessment of writing in independent language learning programs: The value of annotated samples. *Assessing Writing, 10*(3), 174-191.

Brown, J. D. (1998). *New Ways of Classroom Assessment. New Ways in TESOL Series II. Innovative Classroom Techniques*: ERIC.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*: Cambridge University Press.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-31.

Chen, Y. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235-262.

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. doi: 10.1080/03075079712331381064

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

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129-142.

Dickinson, L. (1987). *Self-instruction in language learning* (Vol. 3). Cambridge University Press Cambridge.

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. doi: 10.1080/03075079912331379935

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.

Falchikov, N. (1986). Product comparisonsand process benefits of collabarative peer group and self assessments. *Assessment & Evaluation in Higher Education, 11*(2), 146-166. doi: 10.1080/0260293860110206

Falchikov, N. (1995). Peer Feedback Marking: Developing Peer Assessment. *Innovations in Education & Training International, 32*(2), 175-187. doi: 10.1080/1355800950320212

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(3), 395-430.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment: A meta-analysis comparing peer and teacher remarks. *Review of Educational Research, 70*(3), 287-322.

Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal, 39*(1), 69-95.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*(1), 12-20.

Jiliang, C., & Kun, W. (2007). An investigation of scorer reliability of self and peer assessment of EFL writing among Chinese college students. *CELEA Journal, 30*(1), 3-11.

Leach, L. J. (2000). *Self-directed learning: Theory and practice.* University of Technology, Sydney, Australia, unpublished doctoral thesis.

LeBlanc, R., & Painchaud, G. (1985). Self-Assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673-687.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 484-509.

Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer-and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62.

London, M., & Tornow, W. W. (1998). 360-degree feedback: More than a tool. *WW Tornow, M. London, & CCL Associates (Eds.), Maximizing the value of*, 1-8.

Longhurst, N., & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation, 23*(4), 319-330.

MacKenzie, L. (2000). Occupational Therapy Students as Peer Assessors in Viva Examinations. *Assessment & Evaluation in Higher Education, 25*(2), 135-147. doi: 10.1080/713611424

Magin, D., & Helmore, P. (2001). Peer and Teacher Assessments of Oral Presentation Skills: How reliable are they? *Studies in Higher Education, 26*(3), 287-298. doi: 10.1080/03075070120076264

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 075-100.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman London.

Nakamura, Y. (2002). *Teacher Assessment and Peer Assessment in Practice. (Educational Studies 44*). Tokyo, Japan: International Christian University. (ERIC Document Reproduction Service No. ED464483).

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. doi: 10.1080/02602930050025006

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1-13.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing, 19*(2), 109-131.

Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*(1), 51-63. doi: 10.1080/0260293042003243896

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20.

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.

Sambell, K., McDowell, L., & Sambell, A. (Eds.). (2006). *Supporting diverse students: Developing learner autonomy via assessment*. New York: Routledge.

Sullivan, K., & Hall, C. (1997). Introducing Students to Self-assessment. *Assessment & Evaluation in Higher Education, 22*(3), 289-305. doi: 10.1080/0260293970220303

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility *Optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Springer.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20-27. doi: 10.1080/00405840802577569

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. doi: 10.1080/713611428

Yamashiro, A. (1999). *Using structural equation modeling to validate a rating scale.* Paper presented at the 21st Language Testing Research Colloquium, Tsukuba, Japan.

Appendix A: Paragraph Evaluation Sheet

Criterion (sub-skill)	Description and elements					
Arrangement of Ideas and Examples (AIE)+ Communicative Quality (CQ) or Coherence and Cohesion (CC)	1) presentation of ideas, opinions, and information					
	2) aspects of accurate and effective paragraphing					
	3) elaborateness of details					
	4) use of different and complex ideas and efficient arrangement					
	5) keeping the focus on the main theme of the prompt					
	6) understanding the tone and genre of the prompt					
	7) demonstration of cultural competence					
	8) range, accuracy, and appropriacy of coherence-makers (transitional words and/or phrases)					
	9) using logical pronouns and conjunctions to connect ideas and/or sentences					
	10) logical sequencing of ideas by use of transitional words					
	11) the strength of conceptual and referential linkage of sentences/ideas					
Sentence Structure Vocabulary (SSV)	1) using appropriate, topic-related and correct vocabulary (adjectives, nouns, verbs, prepositions, articles, etc.), idioms, expressions, and collocations					
	2) correct spelling, punctuation, and capitalization (the density and communicative effect of errors in spelling and the density and communicative effect of errors in word formation (Shaw & Taylor, 2008, p. 44))					
	3) appropriate and correct syntax (accurate use of verb tenses and independent and subordinate clauses)					
	4) avoiding use of sentence fragments and fused sentences					
	5) appropriate and accurate use of synonyms and antonyms					