Iranian Journal of Applied Linguistics (IJAL), Vol.18, No.2, September 2015, 77-107

Rater Errors among Peer-Assessors: Applying the Many-Facet Rasch Measurement Model

Rajab Esfandiari*

Assistant Professor of Applied Linguistics,

Imam Khomeini International University, Qazvin, Iran

Received 10 May 2015; revised 23 July 2015; accepted 15 August 2015

Abstract

In this study, the researcher used the many-facet Rasch measurement model (MFRM) to detect two pervasive rater errors among peer-assessors rating EFL essays. The researcher also compared the ratings of peer-assessors to those of teacher assessors to gain a clearer understanding of the ratings of peer-assessors. To that end, the researcher used a fully crossed design in which all peer-assessors rated all the essays MA students enrolled in two Advanced Writing classes in two private universities in Iran wrote. The peer-assessors used a 6-point analytic rating scale to evaluate the essays on 15 assessment criteria. The results of Facets analyses showed that, as a group, peer-assessors did not show central tendency effect and halo effect; however, individual peer-assessors showed varying degrees of central tendency effect and halo effect. Further, the ratings of peer-assessors and those of teacher assessors were not statistically significantly different.

Keywords: Peer-assessment; Rater effects; Rating; Many-facet Rasch measurement model

*Corresponding author: English Language Department, Faculty of Humanities, Imam Khomeini International University, Nourozian Blvd., Qazvin, Iran.

Email address: esfandiari@hum.ikiu.ac.ir

[DOI: 10.18869/acadpub.ijal.18.2.77]

Introduction

With the introduction, development, and acceptance of alternative assessment in education in general (Herman, Aschbacher, & Winters, 1992) and in language education in particular (Brown & Hudson, 1998) as a popular assessment procedure to evaluate language learners' linguistic abilities (Richards & Schmidt, 2002; Ross, 2005), language educators can benefit from the advantages which the various types of alternative assessment procedures may offer. Focusing on the variety of alternative assessment procedures, Huerta-Macías (1995) as well as Brown and Hudson (1998), listed different types of alternative assessment including, but not limited to, checklists, journals, portfolios, self-assessments, and peer-assessments.

Highly effective and pedagogically useful, peer-assessment has been developed rapidly over the past two decades (Zhao, 2014). The benefits obtained from empirical studies attest to the usefulness of peer-assessment in language education (Birjandi & Hadidi Tamjid, 2012; Sadeghi & Abolfazli Khonbi, 2015; Saito, 2008; Topping, 1998; Zhao & Gallant, 2012). Peer-assessment aids students in reflecting on their learning by observing other students' performance (Falchikov, 1986; Gielen, Dochy, & Onghena, 2011; Nulty, 2010; Somervell, 1993; Vickerman, 2009), generates positive attitudes in students (Haaga, 1993; Murakami, Valvona, & Broudy, 2012; Saito & Fujita, 2004), develops a sense of shared responsibility among students (Saito, 2008), and increases a higher level of cognitive thinking (Cheng & Warren, 2005; Davis, 2009).

One form of peer-assessment is peer rating in which language learners' rate their classmates' performance based on a set of assessment criteria (Mostert & snowball, 2012; Pope, 2001; Weaver & Esposto, 2012). Peer-assessors may not produce accurate ratings; therefore, they may assign their peers higher or lower ratings than expected. These unexpected ratings stem from construct-irrelevant factors collectively known as rater errors (Engelhard, 1994) and result from the peer-assessors' inaccuracy, affecting the reliability and validity of the ratings (Eckes, 2012).

Rater errors threaten the fairness of the ratings awarded to students. For example, Eckes (2009) and Schaefer (2008) asserted that subjectivity of raters' judgments could overestimate, or underestimate, the reliability of the ratings. Researchers have categorized rater errors into five main groups: severity and

[DOI: 10.18869/acadpub.ijal.18.2.77]

leniency, bias, restriction of range, halo effect, and central tendency effect (Engelhard, 1994; Myford & Wolfe, 2003). In the following section, first I fully describe peer-assessment as well as its potential values and summarize empirical studies conducted on peer-assessment; next, I explain rater errors generally neglected in peer-assessment studies; and finally, I discuss the many-facet Rasch measurement model, the methodological framework used in this study. This model is promising in rater-mediated assessments. As McNamara (2011) commented, "the advent of multi-faceted Rasch models … constituted a quantum leap in our capacity to investigate and …. to estimate the impact of various aspects or facets of the assessment setting" (pp. 436). Schaefer (2008) concluded that the many-facet Rasch measurement approach "can …. contribute to the fairness and accuracy of performance-based writing assessment" (p. 490).

Review of the Related Literature

Peer-assessment

Researchers define peer-assessment differently. Simply stated, peer-assessment refers to the evaluation of peers' performance (Brown & Hudson, 1998). In Topping's words (2010), 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" (p.62). Similarly, Van Gennip, Segers, and Tillema (2009) defined peer-assessment "fundamentally an interpersonal process in which a performance grade exchange is being established and in which the core activity is feedback given to and received from others …." (p. 42). Loddington (2008) viewed peer-assessment as the process during which language learners assess the performance of their peers on certain specific tasks, using some pre-established criteria. As these definitions demonstrate, researchers use peer-assessment for both formative (learning) and summative (rating) purposes.

Peer-assessment can have a number of advantages. Firstly, peer-assessment helps language learners to develop a sense of ownership, take responsibility for their own learning, and improve their motivation (Loddington, 2008; Topping, 2010; Topping, Smith, Swanson & Elliot, 2000; Wasson & Vold, 2011). Secondly, peer-assessment provides language learners with opportunities to hone their skills (Ballantyne, Hughes & Mylonas, 2002; Tsui & Ng, 2000), establish interpersonal

[DOI: 10.18869/acadpub.ijal.18.2.77]

relationship (Cheng & Warren, 2005; Earl, 1986; Noonan & Duncan, 2005), share ideas (Pope, 2001), and reflect on learning (Falchikov, 1986). Thirdly, peer-assessment has proved effective in self-critical awareness (Falchikov, 2003; Searby & Ewers, 1997). Finally, as a possible alternative to teacher assessment, peer-assessment may reduce teachers' workload (Fry 1990; Fukazawa, 2010; Pope 2001).

Peer-assessment can also have a number of disadvantages. One serious problem with peer-assessment has to do with its utility because it is "a time-consuming process …. lecturers still need to moderate the students' assessment" (Searby & Ewers, 1997, p. 382). A second problem relates to students' assessment capability. Students may not feel qualified enough, or may not have enough self-confidence to assess their peers (Ballantyne, Hughes & Mylonas, 2002; Orsmond & Merry, 1996; Searby & Ewers, 1997; van Zundert, Sluijs-mans, & van Merrie¨nboer, 2010). A third disadvantage of peer-assessment is that it can be intellectually challenging, and students may not feel comfortable while assessing their classmates (McGarr & Clifford, 2013; Topping et al., 2000).

Some researchers have examined the effect of training on peer-assessment (e.g., Asikainen et al., 2013; Halinen et al., 2013). The findings have shown that a lengthy period of training can affect peer evaluation positively (Stanley, 1992), training can provide students with more feedback and prompt them to interact with each other (Zhu, 1995), training will result in developing writing skill, building more confidence, and using more metacognitive strategies (Min, 2005), peer-assessor training can make peer-assessors more consistent, leading to fewer misfitting peer-assessors (Saito, 2008), and when peer-assessors receive training, they tend to provide more valid ratings (Liu & Li, 2014).

Some other researchers have investigated the effect of psychological traits on peer-assessment. Alfallay (2005), for example, attempted to establish the relationship between some selected psychological and personality traits, including motivation types, self-esteem, anxiety, motivational intensity, achievement, and the accuracy of peer-assessment. Alfallay found that (a) compared to self-assessors and teacher assessors, students tended to overrate the performance of their peers more than their own, (b) compared with teacher assessors, the most accurate peer-assessors were those of higher achiever groups and the least accurate were those of low classroom anxiety groups, with students of high classroom anxiety, integrative

[DOI: 10.18869/acadpub.ijal.18.2.77]

motivation, low motivational intensity, low self-esteem, high motivational intensity, low achievement, instrumental orientation, and high self-esteem assessing their peers moderately, and (c) compared with self-assessors, learners with high self-esteem were the most accurate, those with instrumental motivation, low motivation intensity, low achievers, and integrative orientation assessing their peers less accurately. Following these findings, Alfallay concluded that "peer ratings appear to be more accurate and reliable than some previous studies have suggested" (p. 419).

Another strand of research has focused on psychometric properties of peer-assessment. Topping (1998), Asikainen, Virtanen, Postareff, and Pekka Heino (2014), and Harris and Brown (2013) found that peer-assessment is a reliable and valid method of improving students' learning. Similarly, Topping, Smith, Swanson, and Elliot (2000) reported on the adequacy of the peer-assessment, noting that reliability and validity of peer-assessment may depend on several contextual factors. MacKenzi (2000) concluded that peer-assessment had high and satisfactory reliability. Chang, Tseng, Chou, and Chen (2011) and Zhao and Gallant (2012) demonstrated that peer-assessors' ratings seem to be as valid as those of teacher assessors. Finally, Lin, Liu, and Yuan (2001) supported the use of peer-assessment for a web-based context, concluding that peer-assessment had higher validity compared to self-ratings in the same web-based context.

In sharp contrast to the studies outlined in the previous paragraph, some researchers have claimed that peer assessment may not be as reliable and valid as teacher assessment (Panadero, Romero, & Strijbos, 2013). Cheng and Warren (2005) conducted a study to compare reliability of peer-assessment with that of teacher assessment in rating English language proficiency. They concluded that peer-assessors could not rate their peers as reliably as did teacher assessors. Focusing on formative peer-assessment, Falchikov (2003) suggested that peer-assessors be supplied with extensive training and instruction to act as reliably as desired.

Past research has also demonstrated the generally positive attitudes toward peer-assessment (Garfield, 1999; McGarr & Clifford, 2013; Mok, 2010). Lladó, et al. (2013), for example, examined the attitudes and perception of students toward using peer-assessment. The findings of their study revealed that (a) students

[DOI: 10.18869/acadpub.ijal.18.2.77]

expressed positive attitudes toward peer-assessment, (b) peer-assessment created a greater sense of responsibility among students and helped them to learn from their mistakes, and (c) students found peer-assessment more motivating.

Finally, researchers in Iran (Birjandi & HadidiTamjid, 2012; Birjandi & Siyyari, 2010; Jafarpour & Yamini, 1995; Sadeghi & Abolfazli Khonbi, 2015) have studied the potential effects of peer-assessment. They have reported on the efficacy of peer-assessment on the quality of writing performance, the positive effects of peer-assessment on learning and language learners' positive attitudes towards peer-assessment, and the positive effect of training on the accuracy of the peer-ratings.

Central tendency effect and halo effect as two pervasive rater errors using the many-facet Rasch measurement model

Rater errors generally refer to the errors the raters produce when rating a product. Rater errors, in Scullen, Mount, and Goff's (2000) words, is a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957). Myford and Wolfe (2003) identified five main types of rater effects: severity/leniency, halo effect, central tendency effect, bias, and randomness. In the present study, we mainly focus on halo effect and central tendency effect, also known as centrality. In the following four paragraphs, we try to define these two errors and report on the very few empirical studies.

Halo effect refers to the tendency to generalize one aspect of an individual's personality to all other aspects of his or her personality. Thorndike (1920) coined the term and defined it "as a marked tendency to think of the person in general as either good or rather inferior and to color the judgments of the qualities by their general feeling" (p. 25). In like manner, Ary, Jacobs, Sorensen (2010) defined halo effect as "a generalized impression of the subject to influence the rating given on very specific aspects of behavior" (p. 215). In studies in which the many-facet Rasch measurement model is used, halo effect occurs when raters fail to distinguish between conceptually distinct and independent aspects of ratees' performance and assign them similar ratings across those traits. As Myford and Wolfe (2004) succinctly put it, halo effect refers to "a rater's tendency to assign ratees similar ratings on conceptually distinct traits" (p. 209).

[DOI: 10.18869/acadpub.ijal.18.2.77]

Using the many-facet Rasch measurement model, Engelhard (1994) tried to detect halo effect among the ratings of 15 experienced native-English speakers who rated 264 compositions. The results of Facets analysis showed that two raters failed to distinguish between two out of four categories on a 4-point rating scale, an indication of halo effect. In a study in Japan, Kozaki (2004) used four bilingual judges to rate the translations of Japanese medical students on a 4-point rating scale on seven assessment criteria. The findings illustrated that the two judges assigned unexpectedly harsh ratings to the weakest examinee and unexpectedly lenient ratings to the most able examinee. These unexpected ratings were suggestive of halo effect, "which judges carry over the impression of competence …. creating non-independence of assessment categories …. grammar or vocabulary or both (Kozaki, 2004, pp. 21-22). In Australia, Knock, Read and Randow (2007) found that raters could show some signs of halo effect even after they were trained. In Iran, Farrokhi and Esfandiari (2011) used 188 undergraduate Iranian English majors to rate the essays their classmates had written. The peer-assessors used a 6-point analytic rating scale to assess the essays on 15 assessment criteria. The results of Facets analysis showed that the individual peer-assessors might display halo effect.

Central tendency effect is the assignment of ratings tightly clustered around the midpoints of a scale. Wolfe (2004) was very specific when he explicitly stated that "raters who tend to assign fewer scores at both high and low ends of the rating scale are said to exhibit *centrality*. This results in a concentration of assigned ratings in the middle of the rating scale" (p. 39). More specifically, Myford and Wolfe (2003) noted that, within the context of many-facet Rasch measurement model, "the rater exhibiting this effect overuses the middle category of a rating scale while avoiding the extreme categories" (p. 396). In an empirical study to detect central tendency effect using the many-facet Rasch measurement model among peer-assessors, Farrokhi, Esfandiari, and Vaez (2011) employed 188 peer-assessors to rate the five-paragraph essays their classmates had written. Peer-assessors used a 6-point analytic rating scale to assess the essay on 15 assessment criteria. The Results of Facets analysis illuminated that individual peer-assessors did not exhibit any sign of central tendency effect either at group level or at individual level.

[DOI: 10.18869/acadpub.ijal.18.2.77]

The many-facet Rasch measurement model

The many-facet Rasch measurement model belongs to the family of Rasch models (Eckes, 2011). Linacre (1989) developed this model to account for examinations that include subjective judgments. It was designed to provide a fine-tuned analysis of multiple facets that potentially affect the assessment outcomes where subjective ratings are to be awarded to performances. The many-facet Rasch measurement model is an extension of the Master's (1982) partial credit model that makes possible the analysis of data from assessments that have more than the traditional two facets associated with multiple-choice tests (Myford, 2002). The many-facet Rasch measurement model is essentially an additive linear model that transforms the ordered observations to a logit scale, making it possible to calibrate all parameters on the same equal-interval scale (Bond & Fox, 2015). The logistic transformation of ratios of successive category probabilities (log odds) can be viewed as the dependent variable with various facets, such as peer-assessors, assessment criteria, and students' essays as independent variables that influence these log odds.

The many-facet Rasch measurement model provides three pieces of diagnostic information. It provides a statistical framework, enabling researchers to analyze rating data and summarize the overall ratings in terms of group-level effects for facets such as peer-assessors, assessment criteria, and students' essays (Myford & Wolf, 2003). It also allows researchers to examine individual-level effects of the various elements of facets (i.e., how each individual peer-assessor, student's essay, and assessment criterion functions in an assessment setting) (Linacre, 2004). The many-facet Rasch measurement model supplies the researchers with interactions between various facets so that they can examine patterns in the ratings of assessors (McNamara, 1996).

The many-facet Rasch measurement model has some assumptions and requirements. It is unidimensional, implying all assessment criteria on a rating scale should measure the same single underlying variable, or construct (Eckes, 2009). The many-facet Rasch measurement model is invariant, implying that each facet can be separated out and estimated independently of other facets to determine how various facets are functioning as intended. Students' essay measures are invariant across different sets of assessment criteria, or assessors, and assessment criteria, or assessor measures, are invariant across different sets of students' essays.

[DOI: 10.18869/acadpub.ijal.18.2.77]

In other words, the invariance properties of the Facets model are only achieved when there is good model-data fit. As Linacre (1994) succinctly put it, "parameters—person ability, item difficulty, and judge severity—should be independent of each other" (p. 42). The final assumption of the many-facet Rasch measurement is that peer-assessors, students' essays, and assessment criteria could be uniquely ordered, respectively. When this unique ordering is met, and the data fit the model, as Engelhard (2002) explained, the following characteristics can be attained: (a) separability of parameters with sufficient statistics for estimating these parameters, (b) invariant estimates of assessor severity and criteria difficulty, and (c) equal-interval scales for the measures.

Research questions

Careful examination of the findings of the studies reported in previous sections reveals the following points. Studies on peer-assessment have addressed training effects, efficacy of peer-assessment as a learning tool, reliability, validity, and utility of peer-assessment, students' attitudes to peer-assessment, and psychological traits on peer-assessment. Researchers have also examined rater errors among peer-assessors mostly in L1 settings, restricting themselves to severity/leniency and inconsistency.

Very few researchers have investigated peer-assessment as a measurement tool. Very few researchers have also addressed central tendency effect and halo effect peer-assessors produce in EFL settings. The only relevant studies are those of Farrokhi and Esfandiari (2011) and Farrokhi, Esfandiari, and Vaez (2011). However, these two studies used partial designs in which peer-assessors rated only a single essay. Such incomplete designs increase the amount of measurement error, making it difficult to separate the net effects of the construct. Using peer-assessors for rating purposes merits further investigation, as it is essential that more studies use fully crossed designs in which peer-assessors rate more essays. As such, with the help of the many-facet Rasch measurement model, this study used a fully crossed design to detect central tendency effect and halo effect to shed light on the utility of peer-assessment as a measurement tool. In this study, the ratings of peer-assessors were also compared to those of teacher assessors.

To address these points, the following research questions were formulated:

[DOI: 10.18869/acadpub.ijal.18.2.77]

1. To what extent do peer-assessors produce halo effect when rating EFL essays?
2. To what extent do peer-assessors produce central tendency effect when rating EFL essays?
3. Do peer-assessors and teacher assessors differ on their ratings?

Methodology

Participants

Initially, 60 peer-assessors participated in this study, but 11 peer-assessors were left out of the study based on a placement test specifically given to screen only high proficient peer-assessors for rating. Finally, 39 peer-assessors rated all 20 essays in the study. The peer-assessors were graduate (MA) Iranian students enrolled in Advanced Writing classes, majoring in English Language Teaching in two private universities in Iran. They ranged in age from 21 to 42, with an average age of 27.35. Of 39 peer assessors, 33 (84.6%) were female and six (15.4%) were male. Thirty seven (94.9%) were native Farsi-speakers, one (2.6%) was native Turkish-speaker, and one did not indicate his/her mother tongue. Thirty five peer-assessors were first-year students, three were second-year students, and one was a third-year student. Only one assessor had experience living in an English-speaking country. Thirty six assessors (92.3%) had studied English before entering the university, and the number of years they had studied English ranged from 7 to 30 years.

In addition to peer-assessors, two experienced English language teachers with extensive rating experience participated in this study. One of them was male, and the other one was female. Both of them were native-Farsi speakers, and they were 26 and 38 years old. The male teacher had a PhD in English Language Teaching, and the female teacher was an MA holder in English Language teaching. They had taught Writing from 3 to 12 years. None of them had the experience of living in an English-speaking county.

The instrument

The researcher decided to use an analytic rating scale for the following three reasons. Analytic rating scales provide diagnostically useful information (Knock, 2014; Weigle, 2002), they are more reliable (Hamp-Lyons, 1991), and they tend to be more useful in rater training, as "inexperienced raters can more easily understand and apply the criteria" (Weigle, 2002, p. 120).

[DOI: 10.18869/acadpub.ijal.18.2.77]

The instrument the researcher used contained 15 assessment criteria: (1) substance, (2) thesis development (3) topic relevance, (4) introduction, (5) coherent support, (6) conclusion, (7) logical sequencing, (8) range, (9) word choice, (10) word form, (11) sentence variety, (12) overall grammar, (13) spelling, (14) essay format, and (15) punctuation, capitalization, and typing. The scale categories were very poor (1), poor (2), fair (3), good (4), very good (5), and excellent (6). The researcher borrowed the scale from Esfandiari and Myford (2013). Since in the present study, typed essays were used for rating, assessment criterion 15 was changed from punctuation, capitalization, and handwriting to punctuation, capitalization, and typing with no difference in proper functioning of the scale. Interested readers are referred to Esfandiari and Myford for more information on psychometric properties and other issues relating to the construction of the scale.

Rater training

The researcher conducted a two-hour training session in order to provide the peer-assessors with the steps and procedures involved in rating the essays. The researcher explained the purposes of peer-assessment to peer-assessors. After explaining the purposes, the researcher gave peer-assessors an essay which had already been rated by an expert teacher and a set of guidelines in Farsi explaining the assessment criteria in the rating scale in detail.

The researcher clarified the assessment criteria while drawing peer-assessors' attention toward the corrected essay so that they could analyze the rated essay based on the guidelines. The researcher asked peer-assessors to rate their peers' essays using different categories of the rating scale. The researcher then gave another essay to the peer-assessors to rate in the classroom. While the peer-assessors were busy rating the essay, the researcher checked the peer-assessors' ratings to ensure they understood the process. The researcher was monitoring the peer-assessors while they were rating their peers' essays to make sure they were rating as accurately as possible. Peer-assessors were given a second essay to rate in the class.

After the researcher ensured that the peer-assessors were familiar with the rating procedures, he assigned each peer-assessor 20 essays to rate at home. To rate as accurately as possible and encourage peer-assessors, the researcher told them the

[DOI: 10.18869/acadpub.ijal.18.2.77]

assignment (peer-assessing the essays) was worth 20% of the final course grade. The essay writers' names were removed so that peer-assessors could not figure out whose essay they would rate. The researcher asked them to return the rated essays at the final exam session.

Data collection methods

The researcher collected the data from two intact classes at two private universities in Iran over a span of six months in Winter 2014: Islamic Azad University, Qazvin branch and Islamic Azad University, sciences and research campus in Qazvin, Iran. Students were in two Advanced Writing classes. They learned different features of writing a five-paragraph essay such as thesis statement, motivator, blueprint, topic sentence, reworded thesis statement, and clincher in these classes.

As to their assignment, the researcher asked them to write an essay on the following topic: as more and more students enter universities, academic qualifications become developed. To get ahead in many professions, more than one degree is now required and in the future, it is likely that people will take a number of degree courses even before starting work. Do you agree or disagree? The students had to type the essay and submit it to the instructor in the class.

After the students returned their assignments to the researcher, he randomly chose 20 essays out of their five-paragraph essays. Following the data collection, the rater training in which the students were trained regarding how to rate the essays was held.

Data analysis procedures

To analyze the rating data from this study, the researcher employed Facets, the Rasch-based rating scale analysis computer software program, version 3.71.4 (Linacre, 2014a). Facets calibrated peer-assessors, assessment criteria, and students' essays as independent variables, or facets, simultaneously, but independently of each other, so that all facets were positioned on the same scale, creating a single frame of reference for interpretation of results from the analysis. To make the analysis more powerful, the researcher used a fully crossed design in which all 39 peer-assessors assessed all 20 students' essays on all 15 assessment criteria. Using a fully crossed design allowed the researcher to create a judging

[DOI: 10.18869/acadpub.ijal.18.2.77]

plan in which each element of each facet was directly linked to that of other facets. The fully crossed design had the added benefit of enabling the researcher to create a common frame of reference in interpreting the results.

Results

Variable map

The researcher decided to present the results of the study around the research questions. However, before turning to the research questions, the researcher provides the readers with the Facets map (Figure 1), the single most informative piece of information, displaying all facets of analysis, summarizing key information about each facet, and enabling us to view all facets of analysis at one time.

The first column of figure 1 shows the logit scale, an equal-interval scale which ranges from 2 to -2. The second column shows peer-assessors. More severe peer-assessors appear at the top of the column, while more lenient peer-assessors appear lower in the column. Peer-assessors 31 and 35 were the most severe assessors, while peer-assessors 25 and 33 were the most lenient assessors, with the rest of the peer-assessors appearing midway between these two groups of peer-assessors. Column three displays students' essays. Essays higher on the scale receive lower ratings, but those lower on the scale receive higher ratings. The fourth column shows the assessment criteria. Peer-assessors severely rated assessment criteria higher on the scale, but they rated those lower on the scale leniently. Therefore, assessment criterion 9 was the most difficult one for students to receive low ratings on; by contrast, assessment criterion 13 was the easiest one for the students to receive high ratings on. The last column shows the categories of the scale from 1 to 6.

[DOI: 10.18869/acadpub.ijal.18.2.77]

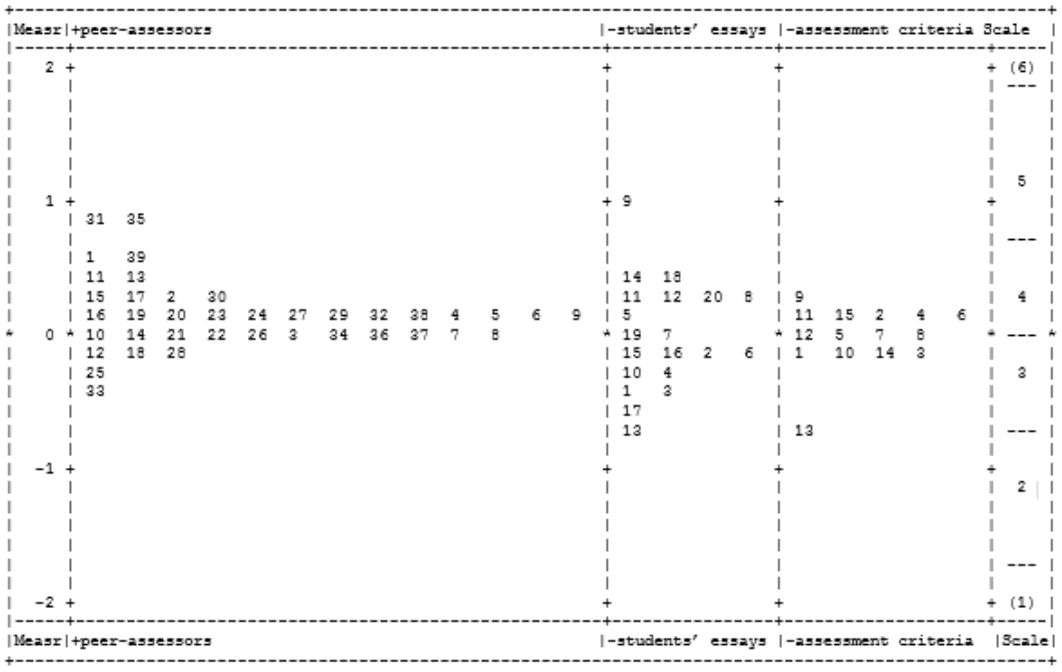

Figure 1:

Variable map showing peer-assessors, students' essays, and assessment criteria

Halo effect among peer-assessors

Facets generates two groups of statistics to detect halo effect: group-level and individual-level statistics. To investigate halo effect at group level, the researcher examined the differences among the difficulty level of the assessment criteria. When the majority of peer-assessors show halo effect, they tend to assign each student similar ratings across conceptually distinct assessment criteria, thereby being unable to distinguish among assessment criteria. Three statistical indicators in Facets are useful indicators if differences in the difficulty of assessment criteria exist.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Assessment criteria separation index: This indicator shows the number of statistically distinct levels of assessment criteria difficulty among the assessment criteria. A low assessment criterion separation index suggests halo effect among the ratings of peer-assessors.

Reliability of peer-assessor separation index: This indicator tells us how well the peer-assessors are able to distinguish among the assessment criteria. A low peer-assessor separation reliability index implies halo effect among the ratings of peer-assessors.

A "fixed-effect" chi-square test of homogeneity: This tests the hypothesis that all assessment criteria share the same level of difficulty, after allowing for measurement error. A non-significant chi-square value suggests halo effect among the ratings of peer-assessors.

Assessment criteria separation index was 11.08, which implies that there were about 11 statistically distinct levels of assessment criteria difficulty in the sample of assessment criteria. Reliability of peer-assessor separation index was .98, suggesting the peer-assessors could reliably distinguish among the assessment criteria. The "fixed-effect" chi-square test of homogeneity was statistically significant **(**$\chi^2(14) = 761.3$, $p<.05$). On a group level, peer-assessors did not show any sign of halo effect.

On an individual level, the researcher followed the approach suggested by Linacre (2014). The following steps were taken to detect halo effect among peer-assessors. First, the researcher anchored all assessment criteria at the same difficulty, which is usually set at 0. Linacre (2014b), for example, has advised us to "anchor all items at the same difficulty, usually 0. Raters who best fit this situation are most likely to be exhibiting halo effect"(p.319). Second, the researcher examined fit indices to detect peer-assessors who fit the model. As Eckes (2011) asserted, "raters whose fit values gravitate toward the expected value would be likely to demonstrate halo (pp. 67-68). Finally, the researcher set .70, above which peer-assessors did not show halo. This cut-off point is arbitrary because neither Linacre nor Eckes defined a cut-off score. Table 1 includes information on those peer-assessors who demonstrated halo effect in this study.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Table 1
Cases of halo effect among peer-assessors

Peer-assessors	Infit	Outfit
15	.47	.46
22	.48	.47
33	.49	.50
25	.58	.60
4	.59	.58
2	.63	.64
30	.63	.63
7	.64	.64
3	.65	.65
23	.68	.68
14	.69	.69

As shown in Table 1, 11 out of 39 peer-assessors seemed to exhibit halo effect based on their relatively great tendency toward overfit, as shown in columns two and three. To ensure these peer-assessors assigned identical ratings, the researcher examined the actual ratings for peer-assessors. As an example, peer-assessor 15 assigned many identical ratings across assessment criteria while rating the twenty essays. For example, this peer-assessor assigned the 17th essay identical ratings of 5 across the first 13th assessment criteria as shown below: 15,17,1-15,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,4. 15 stands for peer-assessor; 17 stands for the essay; 1-15 stands for assessment criteria; and 5,5,5,5,5,5,5,5,5,5,5,5,5,6,4 stands for the ratings this peer-assessor assigned in this study. Such flat patterns suggest halo effect.

Central tendency effect among peer-assessors

Myford and Wolfe (2003) noted that, within the context of many-facet Rasch measurement model, "the rater exhibiting this effect [central tendency] overuses the middle category of a rating scale while avoiding the extreme categories" (p. 396). For the second research question, the researcher used the following group-level statistical indicators. ***A "fixed-effect" chi-square test of homogeneity:*** A significant chi-square value for student writing ability measures indicates that at least two students are statistically significantly different in terms of their writing ability (Myford & Wolfe, 2004). ***Chi-square test of homogeneity*** was statistically

[DOI: 10.18869/acadpub.ijal.18.2.77]

significant ($\chi^2(19) = 2322.0$, $p < .05$). ***Student separation index:*** It is an indicator which shows the number of statistically distinct levels of student writing ability. A low student separation index suggests the presence of central tendency effect (Myford & Wolfe, 2004). ***Student separation index*** was 15.27, suggesting multiple statistically distinct levels of writing ability among students. ***Reliability of peer-assessor separation index:*** It is an indicator showing how well students are separated by their writing ability (Myford & Wolfe, 2004). High reliability suggests that students are well separated according to their level of writing ability with high confidence. ***Reliability of peer-assessor separation index*** was .99, implying separation of students' writing ability. ***Fit indices:*** Assessment criterion fit mean-square indices significantly less than 1 may signal central tendency effect (Knoch, Read, & von Randow, 2007). It is not clear how significantly less than 1 fit mean-square indices should be. In this study, the researcher, however, decided if infit and outfit mean-square values were significantly less than .70, they would contribute to central tendency effect. As shown in Table 2, none of the assessment criteria were less than .70.

Table 2
Assessment criteria fit indices for possible central tendency effect

Assessment criteria	**Infit mean square**	**Outfit mean square**
Substance	.77	.78
Thesis development	1.10	1.09
Topic relevance	.91	.91
Introduction	.95	.95
Coherent support	.83	.83
Conclusion	1.07	1.06
Logical sequencing	.96	.96
Range	.81	.82
Word choice	1.02	1.02
Word form	.92	.92
Sentence variety	1.00	.99
Overall grammar	1.08	1.10
Spelling	1.15	1.20
Essay format	1.25	1.24
Punctuation, capitalization, and typing	1.19	1.21

[DOI: 10.18869/acadpub.ijal.18.2.77]

Finally, the researcher examined how many times as a group the peer-assessors used the scale categories. If peer-assessors overused the inner categories, this may be a sign of central tendency effect. Table 3 shows that there is not a pervasive trend toward central tendency effect on the part of all peer-assessors. The ratings are approximately distributed across scale categories, with nontrivial ratings assigned to lower and upper categories (1 and 6).

Table 3
Use of rating scale categories by peer-assessors

Scale category	***N***	**%**
1	805	7
2	1649	14
3	2467	21
4	2997	26
5	2287	20
6	1315	11

On an individual level, the researcher examined peer-assessor infit and outfit mean-square indices for possible central tendency effect. Following Linacre (2002) and Lunz, Stahl, and Wright (1994), in this study, the lower limit .50 and the upper limit 2 were used because, as Wright and Linacre (1994) explained, such limits could be "productive for measurement" (p. 370).

Table 4
Peer-assessors' fit indices for possible central tendency effect

Peer-assessors	**Infit mean-square**	**Outfit mean-square**
1	1.33	1.45
2	.68	.72
3	.72	.71
4	.60	.59
5	.84	.85
6	.89	.88
7	.68	.70
8	.95	.94
9	1.10	1.10
10	1.23	1.22

[DOI: 10.18869/acadpub.ijal.18.2.77]

11	.92	.91
12	1.35	1.30
13	1.66	1.71
14	.75	.77
15	**.49**	**.48**
16	.81	.81
17	1.09	1.08
18	.94	.94
19	.79	.77
20	1.65	1.63
21	1.01	1.01
22	**.47**	**.46**
23	.73	.73
24	1.13	1.14
25	.64	.65
26	.58	.59
27	1.38	1.38
28	.99	.98
29	1.06	1.05
30	.69	.71
31	.85	.86
32	1.34	1.32
33	.54	.55
34	1.37	1.37
35	.76	.82
36	1.55	1.61
37	1.73	1.74
38	1.39	1.41
39	1.18	1.20

Table 4 shows that peer-assessors 15 and 22 were overfitting, suggesting that these two peer-assessors showed less variability in their ratings. This indicates central tendency effect.

[DOI: 10.18869/acadpub.ijal.18.2.77]

The comparison of the ratings of peer-assessors and those of teacher assessors

To answer the third research question about whether the ratings of peer-assessors and those of teacher assessors differed, a Mann Whitney Test was run. The results of descriptive and inferential statistics are presented in Table 5 and Table 6.

Table 5

Results of descriptive statistics for the third research question

	Assessors	**N**	**Mean Rank**
	Peer-assessors	39	20.67
	Teacher assessors	2	27.50
	Total	41	48.17

According to Table 5, teacher assessors had the higher mean rank (mean rank = 27.50) and the lower mean rank (mean rank = 20.67) belonged to peer-assessors. In order to see whether the differences between the mean ranks between these two groups are statistically significant, the Mann Whitney U test was run. The results are presented in Table 6.

Table 6

Results of inferential statistics for the third research question

	Ratings
Mann-Whitney U	26.000
Wilcoxon W	806.000
Z	-.787
Asymp. Sig.(2-tailed)	.431

A Mann-Whitney U Test did not reveal a statistically significant difference in the ratings of peer-assessors (*mean rank* = 20.67, *n* = 39) and those of teacher assessors (*mean rank* = 27.50, *n* = 8), *U* = 26.000, *z* = -.787, *p* = .431, *r* = 6.40).

Discussion

This study was carried out to investigate how reliably peer-assessors would rate the five-paragraph essays their peers had written. More specifically, the researcher aimed to demonstrate whether peer-assessment could be used for summative assessment. The researcher was also interested in how similarly peer-assessors rated compared to teacher assessors. The researcher employed the many-facet

[DOI: 10.18869/acadpub.ijal.18.2.77]

Rasch measurement model to detect central tendency effect and halo effect among peer-assessors.

The results of Facets analyses showed the following findings. Peer-assessors rated the essays of their peers very reliably, not showing central tendency effect and halo effect when the researcher reviewed the group-level statistical indicators. When the researcher compared their ratings with those of the teacher assessors, the ratings did not reveal any statistically significant differences, suggesting that peer-assessors' ratings were as close as possible to those of teacher assessors. When the researcher reviewed individual-level statistics indicators, 11 out of 39 peer-assessors showed halo effect. Only two peer-assessors, however, exhibited central tendency effect.

On average, peer-assessors were reliable assessors, not showing any misfit, and they rated EFL essays as similarly as did teacher assessors. These findings support some of those from previous studies (Esfandiari & Myford, 2013; Farrokhi & Esfandiari, 2011; Chang, et al., 2011; Nakamura, 2002; Zhao & Gallant (2012). Peer-assessors, as a group in this study, seem to distinguish between conceptually distinct assessment criteria, avoiding assigning similar ratings to students across these criteria. Peer-assessors also seem to differentiate among students' ability levels; thus, they understand the distinctions between the middle categories of the rating scale and do not assign students similar "middle-of-the-road" ratings. Possible explanations for peer-assessors not to misfit on average in this study stem from the assertion that individual differences at group level do not usually surface, especially if there are very few differences, as in this study there were only two misfitting peer-assessors for central tendency effect.

The finding that the ratings of peer-assessors and those of teacher assessors resemble each other also confirm some findings of the previous studies (Li, et al. 2015; Liu & Li, 2014). One possible explanation is the amount of training the peer-assessors received in this study. Peer-assessors in this study received a two-hour training in which they were fully briefed on assessing procedures. For example, Liu and Li (2014) concluded that "the training on rubric-based assessment may enable students to become not only better assessors but also better assesses" (p. 287). A second possible explanation may relate to the instruction peer-assessors received. Peer-assessors received instruction during the semester on principles of five-paragraph essays, including punctuation, expression, format,

[DOI: 10.18869/acadpub.ijal.18.2.77]

organization and development of main and supporting ideas, introductory, central, and concluding paragraphs (for a relatively recent discussion, cf. Spatara, Pennaa, Millsa, Kutijaa & Cookea, 2014)). During the semester, the researcher also supplied peer-assessors with detailed feedback on the essays they wrote at home and submitted to them. Such procedures may have familiarized them with the points they needed to follow in rating their classmates' essays.

On an individual level, the ratings of peer-assessors showed both halo effect and central tendency effect, although peer-assessors showed much fewer cases of central tendency effect. Such findings are not surprising given the very fact that even in studies using highly trained professional raters, both halo effect and central tendency effect have been reported (Engelhard, 1994; Myford & Mislevy, 1995; Knock, Read, & Randow, 2007; Kozaki, 2004; Wolfe, Chiu, & Myford, 1999). In rating situations in which raters are being monitored while rating, they might show central tendency effect as a "play-it-safe" strategy (Wolfe, Chiu, & Myford, 1999).

Why those individual peer-assessors in this study showed both central tendency effect and halo effect is partly because they were not able to make fine distinctions among the rating categories; therefore, they tended to assign either middle categories similar ratings or similar ratings to rating categories. Alternatively, the presence of central tendency effect and halo effect might be attributed to their lack of rating experience. In this study, peer-assessors were L2 MA students in an EFL setting, who did not have any rating experience prior to this study. Peer-assessors' only familiarity with rating was the two-hour rating session. Further, they knew that they were rating their classmates' essays. Therefore, they might have felt sympathetic with their classmates and did not award very high or very low scores.

Detecting rater errors among peer-assessors 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 such errors 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 take those. Students may be provided with rich feedback regarding their ratings so that they will incorporate it in their ratings. Properly implemented in language education, such training program can direct students to problematic areas so that they will avoid showing "friendship bias" as far as possible. A second implication relates to graduate courses. Since the ratings of the peer-assessors and

[DOI: 10.18869/acadpub.ijal.18.2.77]

teacher assessors did not reach statistical significance, university professors may use students' ratings with more confidence in their classes. Closely related to the second implication, the third implication concerns concurrent validity. The closer the ratings of peer-assessors and teacher assessors are, the more concurrently valid they are.

Conclusion

Although peer-assessors' ratings in this study were generally as close as possible to those of teacher assessors, individual peer-assessors showed varying degrees of both central tendency effect and halo effect. Having investigated severity differences among three assessor types, Esfandiari and Myford (2013) concluded that "it is still premature to suggest that teachers use peer-assessors' ratings when they are making summative judgments about the writing ability of students in their courses" (p. 127). This conclusion still seems to remain true as demonstrated in this study; admittedly, so few studies do examine peer-assessment as a measurement tool in language education (for the most recent discussion, see Li, et. al. 2016). If we want to do justice to peer-assessment, we should consider examining the measurement dimension of this alternative assessment procedure. Democratic assessment requires that the power between teachers and students be shared. One form of materializing this dictum is to allow students to exercise rating, which has the added benefit of reducing teacher's workload. As the findings of this study showed, group-level statistics are not very informative because individual differences are concealed. When researchers use the many-facet Rasch measurement model for rating data, they should report individual-level statistics, as they provide researchers with individual cases which may be flagged for further analysis and directive feedback.

This study was a quantitative investigation of two pervasive errors among peer-assessors in an EFL setting. The absence of a qualitative component limits the value the findings may have. Qualitative investigation of peer-assessors' ratings could have provided us with much richer information about why some of them actually showed central tendency effect, or halo effect. It might have been due to peer-assessors' personality traits, their topical knowledge, background information, attitudes, gender, or status. Each one of these factors merits further investigation in future studies in which peer-assessors are to assign ratings.

[DOI: 10.18869/acadpub.ijal.18.2.77]

The following points should also be taken into account when researchers intend to investigate peer-assessors' ratings in future studies. First, to gain more reliable results, researchers should supply peer-assessors with sufficient training on assessment criteria, engage them in active co-construction of assessment criteria, and provide them with directive feedback. Second, peer-assessors should be asked to supply qualitative feedback for essays in addition to ratings since when numerical ratings and qualitative information are combined, the accuracy of peer-assessment may be improved. Third, peer-assessors should volunteer to rate the essays, and peer-assessment should be non-anonymous. Voluntary peer-assessment may motivate students to engage in rating, and non-anonymous rating is likely to lead to taking the rating task more seriously, thereby improving the accuracy of the peer ratings.

Notes on contributor

Rajab Esfandiari is an assistant professor at Imam Khomeini International University in Qazvin, Iran. His areas of interest include teaching and assessing L2 writing, many-faceted Rasch measurement, and L2 classroom assessment.

References

Alfallay, I. (2005).The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System*, *32*(3), 407-425.

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). *Introduction to research in education* (7th ed.). Belmont, CA: Thomson Wadsworth.

Asikainen, H. Virtanen, V. Postareff, L. Heino, P. (2014). The validity and students' experiences of peer assessment in a large introductory class of gene technology. *Studies in Educational Evaluation, 43*(4), 186-196.

Asikainen, H., Parpala, A., Virtanen, V., & Lindblom-Ylänne, S. (2013). The relationship between student learning process, study success and the nature of assessment. A qualitative study. *Studies in Educational Evaluation, 39*(4), 211–217.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*(5), 427–441.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Birjandi, P., & Hadidi Tamjid, P. (2012). The role of self-assessment, peer-assessment, and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education*, *37*(5), 513-533.

Birjandi, P., & Siyyari, M. (2010). Self-assessment and Peer-assessment: A Comparative Study of Their Effect on Writing Performance and Rating Accuracy. *Iranian Journal of Applied Linguistics*, *13*(1), 23- 45.

Bond, T., & C. M. (2015). *Applying the Rasch model fundamental measurement in the human sciences* (3rd edition). New York and London: Routledge.

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

Chang, C-C., Tseng, K-H, Chou, P-N, & Chen, Y-H. (2011). Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education, 51*(1), 1306-1316.

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

Davies, P. (2009). Review and reward within the computerized peer-assessment of essays. *Assessment & Evaluation in Higher Education, 34*(3), 321–33.

Earl, S. E. (1986) Staff and peer assessment—measuring an individual's contribution to group performance, *Assessment and Evaluation in Higher Education*, *11*(1), 60–69.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2011). *Introduction to many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9*(3), 270-292.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31(2),* 93-112.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-Scale Assessment Programs for All Students: Development, Implementation, and Analysis* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates.

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 comparisons and process benefits of collaborative peer group and self-assessments. *Assessment & Evaluation in Higher Education, 11(2),* 146–165.

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

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531-1540.

Farrokhi, F., Esfandiari, R., & Vaez, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal, 15*(11), 76- 83.

Fry, S. A. (1990) Implementation and evaluation of peer marking in higher education, *Assessment and Evaluation in Higher Education, 15*(3), 177-189.

Fukazawa, M. (2010). Validity of peer assessment speech performance. *Annual of Review of English Language Education in Japan, 21*, 181-190.

Gatfield, T. (1999). Examining Student Satisfaction with Group Projects and Peer Assessment. *Assessment and Evaluation in Higher Education, 24*(4), 365-377.

Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education, 36*(2), 137–55.

Haaga, D. A. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*(1), 28–32.

Halinen, K., Ruohoniemi, M., Katajavuori, N., & Virtanen, V. (2013). Life science teachers' discourse on assessment: A valuable insight into the variable conceptions of assessment in higher education. *Journal of Biological Education, 48*(1), 16–22.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwod, NJ: Ablex.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 36*(2), 101–111.

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

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

Jafarpur, A., & Yamini, M. (1995). Do Self-Assessment and Peer-Rating Improve with Training? *RELC Journal, 26*(1), 63-85.

Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(2), 26–43.

Knock, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33*(1), 77-86.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*(1), 1–27.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Kyung Sun Chung, K. S., & Hoi K. Suen, H. K. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(1), 245-264.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17*(4), 420-32.

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

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). Optimizing rating scale effectiveness. In E. V. Smith & R.M. Smith (Eds.*), Introduction to Rasch measurement* (pp. 257–578). Maple Grove, MN: JAM Press.

Linacre, J. M. (2014a). *FACETS (Version 3.71.4) [Computer software]*. Chicago, IL: MESA Press.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Linacre, J. M. (2014b). *A user's guide to FACETS: Rasch-model computer programmes*. Chicago: Winsteps.com.

Liu, X., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education, 39*(3), 275-292.

Lladó, A. P.Soley, L. F., Sansbelló, R. F., Pujolras, G. A.,Planella, J. P.,Roura-Pascual, N., Martínez, J. S., &Moreno, L. M. (2014). Student perceptions of peer assessment: an interdisciplinary study. *Assessment and Evaluation in Higher Education, 39*(5), 592-610.

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

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*(4), 914–925.

MacKenzi, L. (2000). Occupational therapy students as peer assessors in viva examinations. *Assessment & Evaluation in Higher Education, 25*(2), 135-147.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47*(2), 149-174

McGarr, O., & Clifford, A. M. (2013). 'Just enough to make you take it seriously': Exploring students' attitudes towards peer assessment. *Higher Education, 65(*6), 677–693.

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

McNamara, T. F. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435-440.

Min, H. T. (2005). Training students to become successful peer reviewers. *System, 33*(2), 293- 308.

Mok, J. (2010). A case study of students' perceptions of peer assessment in Hong Kong. ELT Journal, *65*(3)230-239.

Mostert, M. & Snowball, J. D. (2012): Where angels fear to tread: online peer-assessment in a large first-year class, *Assessment & Evaluation in Higher Education, 38*(6), 674-686.

[DOI: 10.18869/acadpub.ijal.18.2.77]

Murakami, C., Valvona, C., & Broudy, D. (2012). Turning apathy into activeness in oral communication classes: Regular self- and peer-assessment in a TBLT programme. System, *40*(3), 407-420.

Myford, C. M. (2002). Investigating design Features of descriptive graphic rating scales. *Applied Measurement in Education, 15*(2), 187–215

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system*. Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

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

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical Assessment, Research and Evaluation, 10*(17), 1–8.

Nulty, D. D. (2010). Peer and self-assessment in the first year of university. *Assessment & Evaluation in Higher Education, 36*(5), 493-507.

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

Panadero, E. Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195- 203.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education, 26*(3), 235–246.

Richards, J. K., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. London: Pearson Education.

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

Sadeghi, K., & Abolfazli Khonbi, Z. (2015). Iranian university students' experiences of and attitudes towards alternatives in assessment. *Assessment & Evaluation in Higher Education, 40*(5), 641-665.

[DOI: 10.18869/acadpub.ijal.18.2.77]

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.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465- 493.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

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.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment &Evaluation in Higher Education, 18*(3), 221–233.

Spatar, C., Penna, N., Mills, H., Kutija, V., Cooke, M. (2015). A robust approach for mapping group marks to individual marks using peer assessment. *Assessment & Evaluation in Higher Education, 40*(3), 371-389.

Stanley, J. (1992). Coaching student writers to be effective peer evaluators. *Journal of Second Language Writing, 1*(3), 217–233.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

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

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). New York, NY: Routledge.

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), 146–169.

Tsui, A.B., & Ng. M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2) 147–70.

van Gennip, N. A. E., Segers, M., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal and structural features. *Learning and Instruction, 4* (1), 41–54.

[DOI: 10.18869/acadpub.ijal.18.2.77]

van Zundert, M., Sluijsmans, D. M. A., Könings, K., & van Merriënboer, J. J. G. (2012). The differential effects of task complexity on domain-specific and peer assessment skills. *Educational Psychology, 32*(1), 127–145.

Vickerman, P. (2009). Student perspectives on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*(2), 221-230.

Wasson, B., & Vold, V. (2011). Leveraging new media skills for peer feedback. *The Internet and Higher Education, 15*(4), 255–264.

Weaver, D., & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment & Evaluation in Higher Education, 37*(7), 805–816.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35- 51.

Wolfe, E.W., Chiu, C.W. T., & Myford, C. M. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses*. Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

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

Zhao, H. (2014). Investigating teacher-supported peer assessment for ELT writing. *ELT Journal, 68*(2), 155-168.

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227-235.

Zhu, W. (1995). Effects of training for peer response on students' comments and interaction. *Written Communication, 12*(4), 492–528.

[DOI: 10.18869/acadpub.ijal.18.2.77]