A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Three Types of Assessment

Farahman Farrokhi
University of Tabriz, Iran

Rajab Esfandiari
Imam Khomeini International University, Iran

Edward Schaefer
Ochanomizu University, Japan

Rater effects in performance testing is an area in which much new research is needed (C. M. Myford, personal communication, 23 February, 2010). While previous studies of bias or interaction effect as a component of rater effect have employed experienced teachers as raters (e.g., Schaefer, 2008), the present study uses many-facet Rasch measurement (MFRM) to investigate differential rater effect or rater severity or leniency among three rater types: self-assessor, peer-assessor, and teacher assessor. Essays written in English by 188 Iranian English majors at two state-run universities in Iran were rated both by the students themselves as self-assessors and peer-assessors and by teachers, using a 6-point analytic rating scale. MFRM revealed differing patterns of severity and leniency among the three assessment types. For example, self-assessors and teacher assessors showed the opposite pattern of severity and leniency as compared with peer-assessors when assessing the highest and lowest ability students. This study has implications for the use of peer and self-rating in L2 writing assessment.

JALT Journal, Vol. 34, No. 1, May 2012

評定者効果は今後新たな研究が求められる分野である（C. M. Myford, personal communication, 2010年2月23日）。評定者効果の構成要素の一つであるバイアスや交互作用効果に関する従来の研究は、経験豊富な教員を評定者としたものであるが（Schaefer, 2008）、本研究では多相ラッシュ測定（MFRM）を用い、自己評価、学習者間評価、教員評価の3タイプ間における評定者効果ないし評定者の厳格さ／寛容さを調査した。イランの州立大学2校の188名の英語専攻の学生の書いたエッセイを対象に、自己評価、学習者間評価者として学生たち自身の評価、および教員による6段階分析評定法により評価を行った。MFRMにより、この３つの評価タイプに、厳格さ／寛容さにおいて異なるパターンがあることが明らかになった。例えば、最も能力の高い学生と最も能力の低い学生に対する自己評価と教員評価は、学習者間評価とは反対のパターンの厳格さ／寛容さを示した。以上をふまえL2ライティングの評価における学習者間評価と自己評価使用に関する示唆を与える。

Performance testing involving the use of rating scales has become widespread in the evaluation of second language writing and speaking assessment. With a communicative approach to language teaching, it is felt that this sort of testing gives a fairer reflection of classroom learning and goals than traditional tests. Research into performance testing has focused on student performance, the tests, the scales, and more recently, on raters themselves and what they do when they rate. There has also been interest in the behavior and comparison of different types of raters. One reason for this is that performance testing places an added burden on teachers, since it is more time-consuming than discrete point tests. There has been growing attention paid to the uses of peer-assessment and self-assessment as alternatives or supplements to teacher assessment. If such assessment can be shown to be valid and reliable, it could contribute to lessening the burden on teachers (Fukazawa, 2010).

However, rater judgments do have an element of subjectivity, and this subjectivity has an influence on the reliability and validity of test scores (Eckes, 2009; Lumley, 2005; Schaefer, 2008). Without the implementation of rigorous measurement tools, it is difficult to establish validity or reliability for rater judgments. Two studies that compared rater types but fell somewhat short in this regard are Mahoney (2011) and Yamanishi (2004). In his comparison of teacher and student error evaluations on a dictation quiz, Mahoney found that students tended to evaluate peers' written work more leniently than teachers. Yamanishi's study—in which two groups of raters, high school teachers and university students who were teacher candidates, rated high school students' free compositions—found that while the teachers were consistent in their ratings, the teacher candidates rated somewhat inconsistently. However, these studies relied on raw scores in their analysis and thus lack generalizability. Mahoney acknowledges this when he cau-

tions that in his study comparisons of scores across groups from different classes cannot be made (p. 117).

A promising measurement resource in the investigation of rater behavior in performance testing is many-facet Rasch measurement (MFRM; e.g., Eckes, 2008, 2009; Linacre, 1989/1994), an application of the Rasch model (Rasch, 1960), a logistic latent trait model of probabilities which calibrates the difficulty of test items and the ability of test-takers independently of one another, but places them within a common frame of reference (O'Neill & Lunz, 1996). MFRM expands the basic Rasch model by enabling researchers to add the facet of judge severity to person ability and item difficulty and place them on the same logit (log odds units) scale for comparison. Engelhard (1992) states that MFRM improves the objectivity and fairness of the measurement of writing ability because writing ability may be over- or underestimated through raw scores alone if students of the same ability are rated by raters of differing severity. MFRM adjusts for rater variability and thus provides a more accurate picture of ability. Coniam (2008) observes that "the use of raw scores may substantially disadvantage test takers receiving lower final grades—a situation which in some examination situations may result in failure rather than success on a test" (p. 71). Applying MFRM to data from the Hong Kong Certificate of Education (HKCE) public school examination writing section, Coniam showed that writers of the same ability would get a lower grade if they had a severe rater rather than a lenient rater, thus potentially failing a high-stakes test.

While the majority of published Rasch measurement studies have been conducted in English-speaking countries, Rasch measurement has been attracting increasing attention in Asia, as the 2008 study by Coniam shows. Though it is still not well known in Japan, there have been a number of Rasch studies here as well. Studies that examined peer-assessment of speaking tests with MFRM include Holster (in press) and Fukazawa (2010). Fukazawa used MFRM to investigate the validity of peer-assessment in a Japanese high school setting and concluded that peer-assessment has sufficient validity for assessing speeches in a Japanese high school. However, as in Mahoney (2011), Fukazawa found that student raters rated their peers more leniently than teachers. In a study of peer-assessment of oral presentations given by Japanese university students, Holster used the quality control feature of MFRM known as fit statistics to show that peer raters were highly misfitting, suggesting that they were interpreting the scoring rubric in different ways from teacher raters. He argued that MFRM can be used to provide diagnostic feedback for peer-assessors with idiosyncratic rating patterns, but that

given the high rate of misfit, it is more suited to low-stakes classroom testing rather than high-stakes tests.

Although the studies described thus far examined rater behavior in performance testing, they did not take advantage of another feature of MFRM called bias analysis, which is valuable in studying rater behavior more deeply. As the present study uses bias analysis to investigate rater differences, it is necessary to explain bias analysis in detail.

The bias analysis function of MFRM investigates rater variability in relation to the other facets in the Rasch model. The term bias refers to rater severity or leniency in scoring, and has been defined as "the tendency on the part of raters to consistently provide ratings that are lower or higher than is warranted by student performances" (Engelhard, 1994, p. 98). Wigglesworth (1993) further stated that bias analysis identifies "systematic subpatterns" of behavior occurring from an interaction of a particular rater with particular aspects of the rating situation (p. 309). It can help researchers explore and understand the sources of rater bias, thus contributing to improvements in rater training and rating scale development. In the present study, we use bias analysis to investigate differential rater functioning and rater severity and leniency, in which rater types display favorable or unfavorable inclinations toward either individual students, individual assessment criteria, or items of the rating scale (cf. Du, Wright, & Brown, 1996; Ferne & Rupp, 2007; Knoch, Read, & von Randow, 2007).

Previous Bias Analysis Studies

Bias analysis studies search for unexpected interactions, such as those between rater judgments and test-takers' performance. In one study, Wigglesworth (1993; 1994) looked at rater-item, rater-task, and rater-test type interaction in the speaking portion of the Australian Assessment of Communicative English Skills (*access:*), an English skills test for potential immigrants to Australia. She found significant variation in how raters responded to different test criteria. Some rated grammar more harshly, and others rated it more leniently. Likewise, some raters were stricter on fluency or vocabulary, while others rated these more leniently. Moreover, raters differed from each other in their strictness or leniency towards the different task types. Also in Australia, McNamara (1996), in analyzing the results of the Occupational English Test (OET), found that trained raters were overwhelmingly influenced by candidates' grammatical accuracy, contrary to the communicative spirit of the test, and that the raters themselves were unaware of this. McNamara noted that this study showed the usefulness of MFRM in

revealing underlying patterns in ratings data and fundamental questions of test validity.

Lumley (2005) used MFRM and think-aloud protocols to analyze the writing component of the STEP (Special Test of English Proficiency), another high-stakes test for immigrants to Australia. Initially, MFRM was used to eliminate misfitting raters. Ultimately, four trained raters rated 12 writing samples consisting of two tasks each, for a total of 24 samples, which had been taken for research purposes from a pool of STEP test examination papers. MFRM analyses of these samples found significant differences between raters. Like McNamara (1996), Lumley also found that grammar was the most severely rated category.

In his MFRM study of rater bias patterns in a Japanese EFL setting, Schaefer (2008) employed 40 native English speakers to rate 40 essays written by Japanese university students. Each rater rated all 40 essays, using a 6-point analytic rating scale consisting of five categories (Content, Organization, Style and Quality of Expression, Language Use, Mechanics, and Fluency). The results showed that for 11 of the raters, "if Content and/or Organization were rated severely, then Language Use and/or Mechanics were rated leniently, and vice versa" (p. 465). Schaefer also found that "some raters also rated higher ability writers more severely and lower ability writers more leniently than expected" (p. 465).

Addressing self-assessment, peer-assessment, and teacher assessment, Matsuno (2009) used MFRM with 91 students and four teacher raters to investigate how self- and peer-assessments work in comparison with teacher assessments in actual university writing classes in Japan. He conducted a bias analysis of rater-writer interactions and found that "self-raters tended to assess their own writing more strictly than expected" (p. 91). Moreover, in this study "high-achieving writers did not often rate their peers severely and low-achieving writers did not often rate their peers leniently" (p. 92), but peer-assessors showed "reasoned assessments independent of their own performances" (p. 92). Finally, teacher assessors showed relatively individual bias patterns.

In investigating the phenomenon of rater subjectivity and inconsistency in L2 performance testing, MFRM allows researchers to analyze rater effects at both group and individual levels (Myford & Wolfe, 2004). Bias analysis can identify patterns in ratings unique to individual raters or across raters, and whether these patterns, or combinations of facet interactions, affect the estimation of performance. However, most of these studies have examined individual rater variation. There still do not seem to be many studies that

have investigated the possibility of systematic patterns in rater *type* variation. Eckes (2008) used cluster analysis following an MFRM bias study to identify the existence of rater types, but these types emerged from group scoring profiles, such as Structure Type and Fluency Type. The present study defines type as a preexisting group, that is, self-assessor and peer-assessor (student assessors) and teacher assessor. The only study to our knowledge that has used bias analysis to investigate rater variation in self-assessment, peer-assessment, and teacher assessment (Matsuno, 2009) did not concentrate on rater type patterns at all. Given the paucity of research on systematic bias patterns in rater type, this area warrants further research.

Furthermore, previous bias studies have either dealt with ESL situations as opposed to EFL, or utilized native L1 raters of L2 essays. Negishi (2010) used MFRM to analyze Japanese L1 raters' assessment of the group oral EFL interactions of Japanese secondary and university students (though this was not a bias study), but the other studies reported above all fit this pattern. There is a need for EFL studies that employ nonnative English-speaking raters of EFL essays, since that is the reality of ELT testing in many countries.

The Present Study

In the present study, MFRM was employed to investigate differential rater severity and leniency with nonnative English speaker raters in an EFL situation. We were interested in how three rater types, self-assessor, peer-assessor, and teacher assessor, interact with the assessment criteria or items of the rating scale. Closely related to this, we were also interested in the severity and leniency of rater type toward students. An important implication of this study is the possibility of student peer and self-assessment as an alternative to teacher assessment. If such ratings can be shown to be equivalent, this could be an argument for the use of self- and peer-assessment as a way to reduce teachers' workload (Fukazawa, 2010).

Research Questions

To achieve the goals of the present study, the following research questions are presented.

RQ1: How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to items?

RQ2: How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to students?

Methodology

Participants

The participants were 194 raters, subdivided into student raters and teacher raters. The student raters were 188 undergraduate Iranian English majors enrolled in advanced EFL writing classes in two prestigious state-run universities in two different cities in Iran, specializing in three fields of study: English Literature, Translation Studies, and English Language Teaching. The student raters were labeled either self-assessors or peer-assessors. The teacher assessors were six Iranian teachers of English.

The student raters ranged in age from 18 to 29, with one over 30, and another with unidentified age. One hundred and thirty-one student raters (69.7%) were female and 57 (30.3%) were male. One hundred and five (55.9%) were native-Farsi speakers, 68 (36.2 %) were native-Turkish speakers, 11 (5.6%) were native-Kurdish speakers and the other four (2.1%) were grouped as "Other." Ninety-five (50.5%) were sophomores, 29 (15.4%) were juniors, and 64 (34.0%) were seniors. Only three (1.6%) of them had the experience of living in an English-speaking country. The number of years they had studied English ranged from 1 to 24, and most of them (61.7%) had studied the English language in language institutes before entering the university.

The teacher assessors were all male, ranging in age from 23 to 36. They came from two language backgrounds: four native-Farsi speakers, and two native-Turkish speakers. None of them had the experience of living in an English-speaking country. They had taught writing courses from 1 to 7 years. Three of them were affiliated with a national university, one of them with a private university, and two were classified as "Other." Each had a degree in English: Three were PhD students in ELT, two had MAs in ELT, and one had a BA in English literature.

The Rating Scale

Generally speaking, there are three types of rating scales in language testing: primary trait, analytic, and holistic (Weigle, 2002). For the purposes of the present study, we chose an analytic rating scale, adapted from Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey's (1981) ESL Composition Profile, but differing from it in many aspects (See Appendix).

To develop our rating scale, we also referred to writing textbooks in the literature because we wanted the scale to reflect the structure of a standard five-paragraph essay, so the following three books were consulted as a guide to composing the scale categories: *Composing With Confidence: Writing Ef-*

fective Paragraphs and Essays (Meyers, 2006), *Refining Composition Skills: Grammar and Rhetoric* (Smalley, Ruetten, & Kozyreve, 2000), and *The Practical Writer with Readings* (Bailey & Powell, 2008). The scale contains 15 items (substance, thesis development, topic relevance, introduction, coherent support, conclusion, logical sequencing, range, word choice, word form, sentence variety, overall grammar, spelling, essay format, and punctuation/capitalization/handwriting).

These 15 items were equally weighted. Although Jacobs et al.'s (1981) five category scales were differentially weighted, it is not clear how those weights were determined (Kondo-Brown, 2002). Hamp-Lyons (1991) recommends using focused holistic scoring when different weights are assigned to different categories in a given context. Schaefer (2008) also observes that different weights predetermine the ranking importance.

There is no consensus in the literature on the optimal number of levels or bands, but for this study, we created a 6-point scoring scale for each item, because "this is the most common number of scale steps in college writing tests, and a large number of steps may provide a degree of step separation difficult to achieve as well as placing too great a cognitive burden on raters, while a lower number may not allow for enough variation among the multifaceted elements of writing skills" (Schaefer, 2008, p. 473).

Data Collection

One hundred and eighty-eight 5-paragraph essays were collected over a year and a half from 188 students. The students came from six classes taught by four instructors, and there were a total of eight weekly meetings. Following the mandatory syllabus set by the Ministry of Sciences, Research, and Technology in Iran, the students were taught punctuation, expression, features of a well-written paragraph, and the principles of a one-paragraph and a five-paragraph essay.

On the midterm exam the week after the last meeting, the students were given 90 minutes to write a five-paragraph essay ranging in length from 500 to 700 words on the following topic, chosen from a list of TOEFL TWE (Test of Written English) topics: *In your opinion, what is the best way to choose a marriage partner? Use specific reasons and examples why you think this approach is the best*. All the students were given the same topic in order to control for topic effect.

Following the data collection, a rating session was held with all the student raters. Before the actual rating, there was a 1-hour training session.

Raters were given an essay rating sheet, one rated essay, and guidelines in Farsi explaining the rating scale in detail. They were told to read the rated essay first without paying attention to the corrections made on the essay. When they finished reading the essay, the researcher conducting the session directed their attention to the corrections made on the essay and the way it was rated on the rating essay sheet. The researcher then explained in detail the rating essay sheet and how the scores had been assigned.

After this, they were given a new essay written by one of the students and told to read the essay and rate it according to the guidelines. They were instructed to closely follow the guidelines, and the researcher monitored the rating process and explained any unclear points.

Following the training session, the actual ratings were held, beginning with self-assessment. The students were given a new rating essay sheet, the guidelines, and their own essays to rate. The researcher advised them to rate as accurately as possible. Following self-assessment, they were given their classmates' essays, with names removed, to rate as peer-assessment. The same rating procedure was repeated. The entire training and rating session took about 2 hours.

The same rating procedure was repeated for teacher assessors. Since it was not possible to arrange for a group meeting, the researcher met with the teachers individually, instructed them how to rate, gave them all 188 essays, and asked them to complete and submit them in one month.

Results

The present study employs a fully crossed design in which all raters rate all essays. The data was analyzed with *Facets 3.68.1*, a software program for MFRM (Linacre, 2011). Three facets were specified for this study: students, rater type, and items.

To answer the first research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to items?), a bias analysis between rater type and items was specified in *Facets*. There were 45 bias terms in all. Table 1 shows the cases of significant rater type by items bias.

Table 1. Rater-Type-Items Bias/Interaction Analysis

Rater type	Logit	Items	Logit	Obs. score	Exp. score	Obs-Exp average	Bias size	Model S. E.	*t*-score	Infit MnSq	Outfit MnSq
Self	-.17	2	.08	704	625.2	.55	-.47	.08	-5.56	1.0	1.0
Self	-.17	3	.26	715	584.5	.92	-.77	.09	-8.63	.9	.9
Self	-.17	4	.06	711	628.9	.57	-.50	.09	-5.81	.9	.9
Self	-.17	6	.06	661	628.0	.23	-.18	.08	-2.36	.8	.8
Self	-.17	7	-.32	620	674.3	-.39	.33	.07	4.48	.5	.5
Self	-.17	8	.13	549	619.9	-.49	.32	.07	4.94	.6	.6
Self	-.17	9	.54	460	519.9	-.42	.25	.06	3.90	1.0	1.0
Self	-.17	10	-.22	615	668.3	-.38	.30	.07	4.21	.7	.7
Self	-.17	13	-.59	689	724.1	-.25	.26	.08	3.20	1.1	1.1
Peer	.05	2	.08	617	563.2	.39	-.28	.07	-3.70	.9	.9
Peer	.05	3	.26	621	515.0	.78	-.54	.08	-6.99	1.1	1.1
Peer	.05	7	-.32	594	642.4	-.35	.26	.07	3.71	.6	.6
Peer	.05	10	-.22	589	628.9	-.28	.21	.07	2.95	.7	.7
Peer	.05	15	.16	516	552.1	-.26	.16	.07	2.43	.9	.9
Teacher	.12	2	.08	3902	4034.7	-.13	.08	.02	3.29	1.1	1.1
Teacher	.12	3	.26	3493	3729.5	-.23	.14	.02	5.72	1.4	1.4
Teacher	.12	4	.06	3989	4093.8	-.10	.06	.02	2.60	.9	.9
Teacher	.12	7	-.32	4652	4549.4	.10	-.08	.03	-2.83	1.5	1.5
Teacher	.12	10	-.22	4522	4428.8	.09	-.07	.03	-2.49	.9	.9

Fixed (all = 0) chi-square: 414.6 *df*: 45 significance: .00: $p < .00$

Note. Items: 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, 15 = Punctuation.

As is evident in the table, the standard errors (SEs) are low, and the mean square fit statistics are good, with no cases of misfit. Out of 45 bias terms, only 19 were significant, with *t*-scores either greater than +2 or smaller than -2. Eleven of the significant interactions are positive (showing severity), and eight of the significant interactions are negative (showing leniency). Rater type showed significant bias toward only 10 out of 15 items (items 2, 3, 4, 6, 7, 8, 9, 10, 13, and 15). Self-assessor shows nine significant interactions, teacher assessor shows five, and peer-assessor also shows five. All three rater types had slightly more cases of severe bias than lenient (self-assessor: 5 vs. 4; peer-assessor: 3 vs. 2; and teacher assessor: 3 vs. 2). The item displaying

the highest *t*-scores was item 3, with teacher assessor severely biased at 5.72 and self-assessor leniently biased at -8.63. Figure 1 presents a graphical representation of rater differences. It can be seen that the gap is particularly large between self-assessor and teacher assessor on item 3 (Topic relevance).

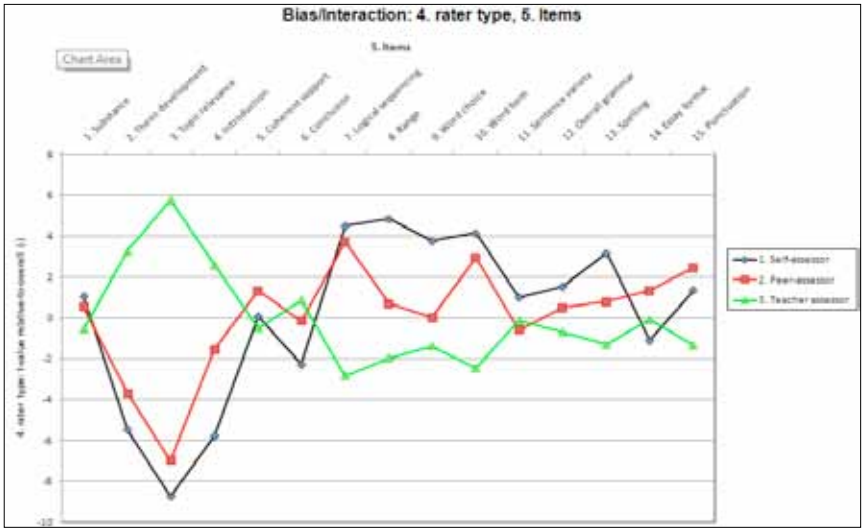

Figure 1. Bias Analysis for Rater Type (Rater-Type-Items Interactions)

Table 2. Frequency of Rater-Type-Items Bias Interactions

Item number	Items	Logit	Self	Peer	Teacher	Total	Lenient/ Severe
2	Thesis development	.09	1L	1L	1S	3	2/1
3	Topic relevance	.29	1L	1L	1S	3	2/1
4	Introduction	.07	1L	0	1S	2	1/1
6	Conclusion	.07	1L	0	0	1	1/0
7	Logical	-.36	1S	1S	1L	3	1/2
8	Range	.14	1S	0	0	1	0/1
9	Word choice	.61	1S	0	0	1	0/1
10	Word form	-.24	1S	1S	1L	3	2/1
13	Spelling	-.65	1S	0	0	1	0/1
15	Punctuation	.18	0	1S	0	1	0/1
Total	10		9	5	5	19	9/10

Note. L = Lenient, S = Severe

As shown in Table 2, which is derived from the information shown in Table 1 and Figure 1, one interesting result is that student raters (self and peer) show the opposite pattern of severity and lenience as compared with the teachers. Students are lenient for items 2, 3, and 4, whereas teachers are severe. However, the opposite is true for items 7 and 10, where students are severe, but teachers are lenient. Both students and teachers have a roughly equal division of severe and lenient interactions with the items, though all three rater types have slightly more severe than lenient bias: 5S/4L for self-assessor, 3S/2L for peer-assessor, and 3S/2L for teacher assessor.

To answer the second research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to students?), another bias/interaction analysis similar to rater-items bias analysis was specified in *Facets* for rater type and students. Table 3 shows the cases of rater-type-by-students bias analysis (due to space limitation, we have included only a small part of the table). The SEs, while low, are much greater than in the rater-items bias analysis, especially for student raters. One cogent reason for the low SEs of teacher assessors is that the total number of teacher ratings across all students far exceeds that of either self-assessors or peer-assessors, because student assessors only rated one student each.

Table 3. Rater-Type-Students Bias/Interaction Analysis

Rater type	Logit	Students	Logit	Obs score	Exp score	Obs-Exp score	Bias size	Model SE	t-score	Infit MnSq	Outfit MnSq
Peer	.05	2	.47	82	63.7	1.22	-1.30	.36	-3.61	.7	.7
Peer	.05	3	.74	83	69.2	.92	-1.16	.38	-3.03	.5	.5
Self	-.17	7	.19	42	62.3	-1.35	.83	.21	4.05	.8	.8
Peer	.05	7	.19	71	57.4	.91	-.65	.24	-2.69	.5	.6
Self	-.17	8	.54	79	69.3	.65	-.69	.31	-2.22	.9	.8
Peer	.05	8	.54	82	65.2	1.12	-1.23	.36	-3.42	1.0	.9
Teacher	.12	8	.54	355	381.5	-.29	.19	.08	2.32	.7	.7
Peer	.05	11	.82	57	70.6	-.91	.64	.20	3.18	.5	.5
Peer	.05	13	.39	42	62.0	-1.33	.82	.21	3.98	.9	.7
Peer	.05	16	.21	57	45.7	.94	-.68	.27	-2.48	.9	.8
Self	-.17	17	1.43	67	75.3	-.59	.72	.26	2.80	.6	.6
Peer	.05	19	.59	79	66.3	.85	-.85	.31	-2.73	1.2	1.1
Peer	.05	24	.54	78	65.2	.85	-.81	.30	-2.71	.8	.8

Fixed (all = 0) chi-square: 1229.7 *df*: 472 significance: $p < .00$

The mean square fit statistics are good, but, unlike rater-type-items bias analysis, there are misfits in rater-type-students bias analysis. A closer inspection of Table 3 shows that out of 19 misfitting rater types, seven belong to self-assessors, 11 to peer-assessors and one to a teacher assessor. It is interesting to note that only one, the teacher assessor, is a case of underfit, and the rest are cases of overfit. Further inspection shows that student logits range from -0.07 to 0.66 for student raters, from -0.07 to 0.56 for self-assessors, and from 0.02 to 0.66 for peer-assessors, and for the teacher assessor the student is a high-ability one with a logit of 1.02. As noted in the literature, underfitting elements show noise, denoting inconsistency (Wigglesworth, 1993, 1994), and overfitting elements show lack of variation, denoting over-predictability (Linacre, 2004). Furthermore, underfitting elements are much more of a problem than overfitting ones (McNamara, 1996). When it comes to deciding how to best deal with either underfit or overfit in an existing data set, Wright, Linacre, Gustafson, and Martin-Lof (1994) claim we should first treat underfits because they force elements to remain below 1. Linacre also recommends retaining overfitting elements in the analysis because, although they do not reveal anything new, at least they tell us something. Besides, due to the lack of students' experience in rating, from a pedagogical point of view it is best to keep overfitting elements to shed light on student rating in classroom settings. Since this is not a validation study to refine an instrument, but rather a study of rater effects, by deleting misfitting elements a good deal of useful information would be lost. Considering the above-mentioned reasons, we decided to let overfitting elements stand as they are. Due to the low number of bias interactions for teacher assessors, we also did not drop the one misfitting teacher assessor.

Out of 472 bias terms, 91 were significant, with *t*-scores either greater than +2 or smaller than -2. Forty-six of the significant interactions are negative (showing leniency), and 45 are positive (showing severity).

Table 4 shows the rater-type-students bias/interaction relationship. To show the relationships, we divided students into four ability groups ranging from -0.35 to 1.45 logits. Across the top of the table is the student logit range, from the lowest ability, -0.35 logits, to the highest, 1.45 logits. Below that is the number of students in each ability group. Finally, the table shows the number of bias interactions for each rater type, divided into severe and lenient ratings.

Table 4. Frequency of Rater-Type-Students Bias Interactions

Student logits	-0.35 to -0.1	0.00 to 0.49	0.50 to 1.00	1.01 to 1.45	Total
n of students	18	105	58	7	188
Severe/Lenient	S/L	S/L	S/L	S/L	S/L
Self	1/2	9/12	7/7	1/0	18/21
Peer	2/1	16/8	5/11	0/1	23/21
Teacher	0/1	2/2	1/1	1/0	4/4
Total	3/4	27/22	13/19	2/1	45/46

Note. L = Lenient, S = Severe

As can be seen in the table, students only have a spread of 1.80 logits, and the majority (164) clusters above the mean (between 0.00 and 1.00). Only 18 students fall below the mean at the lower end of the logit scale, and only seven fall above 1.00 at the upper end of the scale. This could be attributed to the effect of the instruction, in which the students were taught the principles of essay writing, resulting in generally higher ability levels.

The majority of significant bias interactions, 81 out of 91, fall between 0.00 and 1.00 logits, while 10 occur at the extreme ends of the scale. This reflects the fact that the majority of students are clustered just above the mean, with only a relatively small number falling at the lower and upper end of the scale (seven at the lower and three at the upper). Another noteworthy point is that 45 bias interactions are severe and 46 are lenient, showing a roughly equal amount of severe and lenient bias. The third point concerning the table is that rater type shows slightly more lenient bias toward students between 0.00 and 1.00 (41 vs. 40), though again this is roughly equal. The same pattern holds true for rater type bias towards the lowest ability group (4 vs. 3). But when it comes to the highest ability group, the reverse is the case. Rater type seems to be slightly more severe rather than lenient (2 vs. 1). Of course, the low number of bias interactions at the extreme ends makes generalization difficult. There are only seven bias interactions in the lowest ability group and even fewer (only three) in the highest ability groups.

When we compare individual rater types, some interesting patterns emerge. Self-assessor and teacher assessor almost always show more or less the same pattern. For the highest and lowest ability groups, where self-assessor is lenient, teacher assessor is also lenient, and where self-assessor is severe, teacher assessor is also severe. When peer-assessor and teacher assessor are compared, the reverse is true. Where peer-assessor is severe,

teacher assessor is lenient and vice versa. Again this pattern holds true for the lowest and highest ability groups. When self-assessor is compared with peer-assessor, they mostly show the opposite pattern. When self-assessor is lenient, peer-assessor is severe, and when self-assessor is severe, peer-assessor is lenient. Although the small number of cases makes generalization difficult, it seems that self-assessor and teacher assessor ratings resemble each other more than peer-assessor and teacher assessor ratings, or self-assessor and peer-assessor ratings. This finding runs counter to Matsuno (2009) who concluded that "self-assessment was somewhat idiosyncratic and therefore of limited utility as a part of formal assessment" (p. 75).

Overall, self-assessors seem to be the most leniently biased toward students, which is in line with Ross (1998) and Matsuno (2009), who claim that students usually tend to overrate themselves. Peer-assessors are slightly more severely biased toward students, which is in line with Handrahan and Issacs (2001), who also found that peers could be very critical, but teacher assessors show severe and lenient bias in equal measure.

Discussion

The present study used MFRM to investigate bias interactions between three rater types versus first students and then items, and whether these interactions displayed systematic patterns. It further intended to argue for a place for student raters in essay rating in higher education. The findings did discover some recurring patterns. Two types of bias were found: rater type by items and rater type by students. These are explained in detail below.

Student raters (self and peer) show a pattern of severity and lenience toward items that is opposite to that of the teachers. Student raters are lenient for items 2, 3, and 4, whereas teachers are severe. However, the opposite is true for items 7 and 10, where students are severe but teachers are lenient. When we separately analyzed the data for self-assessors, peer-assessors, and teacher assessors, teacher assessors were different from student assessors. The most likely explanation for this is that student assessors were monitored while they were rating the essays while teacher assessors were not. They were rating on their own and they might have had their own interpretations of the criteria, as is quite common in the literature (Lumley, 2005). The monitoring influenced student assessors to have similar rating patterns to each other. This has been shown to result in consistency (see, for instance, Knoch, Read, & von Randow, 2007).

Both self-assessors and teacher assessors show the opposite pattern of severity and leniency as compared with peer-assessors toward the extreme ends of student ability groups. Unlike Kondo-Brown (2002) and Schaefer (2008), who found that raters tended to have more severe or lenient bias toward the extreme ends of ability groups, the present study found that the rater type tended to have more lenient or severe bias patterns toward the midpoints of ability groups, which, as was mentioned in the results section, could be attributed to the instruction students received, making them cluster around the mean, thereby attracting rater type. Like Kondo-Brown's and Schaefer's studies, the present study also found that rater type could show more severe bias toward the highest ability students and more lenient bias toward the lowest ability students. This might be because of the rater type's raised expectations toward the highest ability students or because rater types gave the benefit of the doubt to the lowest ability students, as Schaefer argues, or it might be simply because of the *Facets* program's inability to accurately estimate ability levels at the extreme ends of the continuum, as Kondo-Brown maintains.

Self-assessors tend to have the most severe bias toward items and peer-assessors seem to have the most severe bias toward students. Severity of peer-assessors toward students is mainly because they tend to be critical of their peers and is in line with many previous studies including Handrahan and Issacs (2001). The findings of the present study also corroborate Matsuno (2009) in that peer-assessors in the present study also showed fewer bias patterns toward items, compared to self-assessors. Self-assessors' larger number of bias patterns toward items may be because they did not have a clear understanding of the assessment criteria.

Spelling is the easiest item as scored by rater type. This finding is consonant with Matsuno (2009) and Kondo-Brown (2002) and it is because superficial features like spelling are usually not given in-depth thought (Hamp-Lyons, 2003). It is also in line with Mahoney (2011) who, in the context of error gravity in the Japanese context, asserted that spelling is not as important as other language elements. Word choice is the most difficult item as scored by rater type. This finding runs counter to many previous studies including McNamara (1996), Lumley (2005), Matsuno (2009), and Schaefer (2008). A possible reason could lie in relation to the setting in which the respective studies were done. It seems that different studies in different settings using different raters produce different results concerning the most difficult items and this could be attributed to the perceptions, experiences, and cultural inclinations of raters. McNamara's study was done in an ESL

setting, using highly trained professional raters, and those of Schaefer and Matsuno were done in EFL settings, the former using rather inexperienced native English-speaking raters and the latter using student raters. Another possible interpretation, as it relates to the present discussion and as has been confirmed in previous studies (Saito & Fujita, 2009), might be that raters, especially student raters, are generous in their rating of some items or are unable to differentiate items, hence resulting in inflated marking.

The present study is inconclusive as to whether self- or peer-assessment could be an alternative to teacher assessment in awarding grades on essay writing. There are some inconsistencies. As seen in Table 2, both self-assessors and peer-assessors rate very similarly to each other. In most cases, where self-assessors are lenient, peer-assessors are also lenient and where self-assessors are severe, peer-assessors are also severe. These patterns run counter to teacher assessors who have the opposite pattern. Table 4, however, reveals a different pattern. Here self-assessors and peer-assessors rate mostly differently, and self-assessors rate similarly to teacher assessors. Self-assessors tend to overrate themselves, a finding that is consistent with previous research in which low ability students tend to overrate themselves and high-ability students tend to underrate themselves (Blanche & Merino, 1989; Boud & Falchikov, 1989), which could be attributed to experience (Ross, 1998), subjective points of view such as habits of overestimating of self-ability (Saito & Fujita, 2004), or cultural values of modesty and ego (Brown, 2005). In Iran, evaluation is norm-referenced. When assigning grades, teachers routinely compare a student's work to other students' work (Farhady & Hedayati, 2009). Consequently, when Iranian students rate their own essays, they do not tend to assign ratings that are lower than those that they would assign to their peers' essays. Because students very much appreciate higher ratings, they may be more likely to assign their own essays higher ratings than they actually deserve. Student raters are, however, consistent when it comes to assessment criteria, which is in keeping with Falchikov and Goldfinch (2000), who conclude that when the criteria are explicitly stated and well understood, they lead to more accurate and consistent marking by student raters. The inconsistencies in the present study are partly because the nature of self-assessment and peer-assessment is not yet known and more research is needed to show their efficacy in L2 testing. For example, Saito and Fujita (2004) argue that "lack of research on the characteristics of peer-assessment in EFL writing may inhibit teachers from appreciating the utility of this innovative assessment" (p. 31). Another plausible interpretation for the inconsistency of the results could be the lack

of research using MFRM in this area. As Matsuno (2009) states, "as more researchers use this research technique, we can illuminate a multitude of facets of self and peer assessments" (p. 95). Lack of any meta-analytic study in which findings of other studies are aggregated to arrive at a consensus may well be another reason for the inconsistency.

This study has possible implications for rater training. One hour of training coupled with monitoring in the present study led to more consistency on the part of student raters. Although rater training may not eliminate rater error, it could lead to consistency, especially when it is combined with monitoring. In cases such as this study in which students are involved in rating essays and are going to share rating with teachers in language settings, it is best to provide them with enough training and monitoring. Although the findings in the present study are inconsistent as to the similarity between self, peer, and teacher rating, it was shown that self and teacher ratings were similar to each other, which provides partial evidence for the concurrent validity of self and teacher ratings.

This study is limited in many ways. First, it was purely quantitative. Adding a qualitative component could provide deeper insights into why such findings were obtained. The second limitation relates to the small number of teacher assessors in this study, although the number of teacher assessors was greater than in other studies in which self- and peer-assessments were involved. Most published studies employing self-assessment, peer-assessment, and teacher assessment have used only a very small number of teacher assessors. It would be good for future studies to use a larger number of teacher assessors along with self-assessors and teacher assessors in different settings to see if the results would be the same or different. The third limitation is the small number of essays both self-assessors and peer-assessors rated (one essay each). This is also a further avenue for research, in which future studies could have both self-assessors and peer-assessors rate a larger number of essays because, as was shown in this study, a larger number of ratings could lead to smaller error, which might reduce the number of biases. Finally, due to the small sample size we cannot really make any statements about whether self-assessment or peer-assessment could be a reliable alternative to teacher assessment. The rater bias subpatterns are also based on small differences and need to be interpreted with caution. Researchers should strive to answer this important question in future studies.

Conclusion

Differential rater severity or bias effect as a pervasive rater effect is detrimental if not detected and treated appropriately. In the present study, all three types of rater had bias patterns toward either students or items; furthermore, although these bias patterns were more or less similar, it seems that they were also unique in that each rater type had a particular bias pattern.

Such differences, especially those between students and teachers, seem to be inevitable because they are also manifested in other EFL or ESL settings and confirm previous studies which empirically showed that rater training may reduce rater errors or may make raters self-consistent, but does not necessarily eliminate rater errors (see Knoch, 2011). There are a few methods to help reduce bias when detected. One is to provide rigorous training coupled with monitoring, but the optimal type and amount of training is yet to be shown. In our study, one hour of training led to the consistency of student raters; still, there were many cases of bias, which suggests that it was not enough. Another helpful way is instruction, which might dispense with the need for rater training (Saito, 2008). Instruction might provide raters with clear and explicit assessment criteria or might involve co-creation and negotiation of rating scales with raters. In our study, raters were not provided with such instruction, which might be another reason for bias. Feedback to raters has also proved to be helpful, especially if it is longitudinal (Knoch, 2011). Another reason for the presence of bias in the present study might be lack of feedback.

Coniam (2008) noted that the use of scoring rubrics in performance testing is now common across Asia. The use of peer- and self-raters is also something that is attracting attention in Asia and beyond. MFRM has proven to be useful in researching and testing in this area, and we hope that this study has contributed to the development of this area in EFL settings.

Acknowledgement

We would like to acknowledge the helpful advice of Mike Linacre and Carol Myford regarding the use of the *Facets* computer program to analyze the data.

Farahman Farrokhi is an Associate Professor teaching and supervising MA and PhD students at the University of Tabriz. His areas of interest include classroom management and assessment.

Rajab Esfandiari is an Assistant Professor at Imam Khomeini International University in Qazvin, Iran. His main areas of interest include Rasch measurement.

Edward Schaefer is a Professor in the Graduate School of Humanities and Science at Ochanomizu University. His areas of interest include Rasch measurement and L2 writing instruction and assessment.

References

Bailey, E. P., & Powell, P. A. (2008). *The practical writer with readings*. Blake, VA: Thomson Wadsworth.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning, 39*, 313-340.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529-549.

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

Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *JALT Journal, 30*, 69-84.

Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155-185.

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. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

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

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

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29*, 132-141.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

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

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

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge: Cambridge University Press.

Handrahan, S., & Issacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research and Development, 20*, 53-70.

Holster, T. A. (in press). Many-faceted Rasch analysis of student peer assessment. *Studies in the Humanities*.

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

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing, 28*, 179-200.

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*, 26-43.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese L2 writing performance. *Language Testing, 19*, 3-31.

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

Linacre, J. M. (2004). Optimizing rating scale effectiveness. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

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

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Mahoney, S. (2011). Exploring gaps in teacher and student EFL error evaluation. *JALT Journal, 33*, 107-130.

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

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

Meyers, A. (2006). *Composing with confidence: Writing effective paragraphs and essays.* New York: Pearson Longman.

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

Negishi, J. (2010). Multi-faceted Rasch analysis for the assessment of group oral interaction using CEFR criteria. *Annual Review of English Language Education in Japan, 21*, 111-120.

O'Neill, T. R., & Lunz, M. E. (1996, April). *Examining the invariance of rater and project calibrations using a multi-facet Rasch model*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (rev. ed.). Copenhagen: Danish Institute for Educational Research.

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

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

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

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

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

Smalley, R. L., Ruetten, M. K., & Kozyreve, J. R. (2000). *Refining composition skills: Rhetoric and grammar*. Boston: Heinle & Heinle.

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77-103.

Wright, B. D., Linacre, J. M., Gustafson, J.-E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370. Retrieved from http://rasch.org/rmt/rmt83b.htm

Yamanishi, H. (2004). How are high school students' free compositions evaluated by teachers and teacher candidates? A comparative analysis between analytic and holistic rating scales (in Japanese). *JALT Journal, 26*, 189-207.

Appendix

Essay Rating Sheet

Essay number:

Rater's name:

	Very poor	Poor	Fair	Good	Very good	Excellent
1. Substance	1	2	3	4	5	6
2. Thesis development	1	2	3	4	5	6
3. Topic relevance	1	2	3	4	5	6
4. Introduction	1	2	3	4	5	6
5. Coherent support	1	2	3	4	5	6
6. Conclusion	1	2	3	4	5	6
7. Logical sequencing	1	2	3	4	5	6
8. Range	1	2	3	4	5	6
9. Word choice	1	2	3	4	5	6
10. Word form	1	2	3	4	5	6
11. Sentence variety	1	2	3	4	5	6
12. Overall grammar	1	2	3	4	5	6
13. Spelling	1	2	3	4	5	6
14. Essay format	1	2	3	4	5	6
15. Punctuation/capitalization/handwriting	1	2	3	4	5	6