J. Basic. Appl. Sci. Res., 2(9)8786-8798, 2012

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Self-Assessment, Peer-Assessment, and Teacher Assessment

Farahman Farrokhi[1], Rajab Esfandiari*[2], Edward Schaefer[3]

[1]Associate professor, Department of English Language, University of Tabriz, Tabriz, Iran,
[2]PhD candidate in English Language Teaching, Department of English Language, University of Tabriz, Tabriz, Iran
[3]Full professor, Graduate School of Humanities and Science, Ochanomizu University, Tokyo, Japan

ABSTRACT

Rater effects is an area in which much new research is needed. Rater effects or rater errors have been grouped into five catalogues: rater severity/leniency, halo effect, central tendency effect, randomness or inaccuracy and bias or interaction effect. Bias effect has been extensively studied in the area of L2 testing. However, almost all these studies have employed experienced raters. Not a single study has yet embarked upon investigating rater severity/leniency in self-assessment, peer-assessment, and teacher assessment. The present study is an attempt to investigate differential rater functioning or rater severity/leniency in self-assessment, peer-assessment, and teacher assessment, using a relatively newly employed methodology in language testing, a many-facet Rasch measurement model (MFRM). The study was conducted on one hundred and eighty eight students majoring in English and enrolled in advanced writing classes in two state-run universities. One hundred and eighty eight essays were collected and rated by both students themselves as self-assessors and peer-assessors and teachers using an analytic rating scale on a six-point scale, consisting of fifteen items. The rated data were analyzed using a version of Facets 3.68.0 to answer the research questions. The results of the analysis showed some recurring patterns in three types of assessment. The findings also showed that self-assessors tended to have the most severe bias toward items. Peer-assessors seemed to have the most lenient bias toward students. It was further found that word choice was the most difficult item as scored by rater type, and spelling was the easiest item as scored by rater type. The implications will be discussed in light of rater training and concurrent validity.
Key words: rater type, Many-facet Rasch measurement, rater effects, rating scale.

INTRODUCTION

With the growing use of performance testing in second language writing, in which student writing is assessed by raters using some kind of rating scale, attention has also turned to raters themselves. It is of course desirable that raters rate consistently and objectively, so that ratings reflect student ability rather than factors unrelated to that ability such as rater biases. Researchers recognize that rater judgments have an element of subjectivity, and rater judgments of the same writers often vary (Eckes, 2009; Schaefer, 2008).

A number of researchers have investigated rater behavior through the use of many-facet Rasch measurement (Linacre, 1989/1994). Many-facet Rasch measurement (MFRM) is 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 each other, but places them within a common frame of reference (O'Neill & Lunz, 1996). Measurements of person ability and item difficulty are expressed in units called logits (log odds units). MFRM expands the basic Rasch model by enabling researchers to add the facet of judge severity (or another facet of interest) to person ability and item difficulty and place them on the same logit scale for comparison. It has shown great promise in the area of performance assessment and rating scale validation because it can analyze sources of variation in test scores besides item difficulty or person ability. It is like a 'test' of the raters which measures the rater–scale interaction, just as the person–item interaction is measured (McNamara, 1996). Engelhard (1992) adds that MFRM improves the objectivity and fairness of the measurement of writing ability because writing ability may be over or under estimated 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.

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 sub patterns' 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, however, bias refers to differential rater functioning or rater severity/leniency, which , following Du, Wright, & Brown (1996), Ferne & Rupp (2007), and Knoch, Read, & Von Randow (2007), refers to a situation in which rater types display favorable or unfavorable inclinations toward either individual students or individual assessment criteria or items of the rating scale.

***Corresponding Author**: Rajab Esfandiari, PhD candidate in English Language Teaching, Department of English Language, University of Tabriz, Tabriz, Iran. E-mail: rbesfandiari@gmail.com.

II LITERATURE REVIEW

1 Previous Bias Studies

The idea of searching for unexpected interactions between rater judgments and test takers' performance or other facets in an analysis is central to bias analysis. Many studies to date have found significant differences in rater severity caused by bias interactions. For example, in a study of an Illinois state writing assessment program for middle and high school students in the USA, Du, Wright, & Brown (1996) investigated potential sources of bias based on topic type, ethnic group, and gender. They found that significant rater biases accounted for 11% of the total interactions. The researchers found that raters were biased toward topic type based on ethnic group and gender: "… more raters have bias for or against white students than black students, more raters have bias for or against female students than male students" (p. 15). Du, Wright, and Brown's study is a good illustration of the usefulness of MFRM bias analysis to investigate various facets of interest, such as gender, ethnicity, and topic type.

In the area of L2 studies, 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 rater differences in how candidates responded to different items. Raters displayed an individual pattern of response to the different test criteria (Wigglesworth, 1993, p. 314; 1994, p. 85). Some raters were consistent in their overall ratings, while others were inconsistent. Some rated grammar more harshly, and others rated it more leniently. Likewise, some raters were harsher on fluency or vocabulary, while others rated these more leniently. Moreover, raters differed from each other in their harshness 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. There was a notable contrast between raters' conscious perception of the importance of grammatical accuracy (which was downplayed), and what the MFRM revealed (grammar was important). In other words, there was a difference between what the raters thought they were doing, and what they actually did. McNamara noted that this study showed the usefulness of MFRM in revealing underlying patterns in ratings data and fundamental questions of test validity (1996, p. 216).

Lumley (2002, 2005) used MFRM and think-aloud protocols to analyze the writing component of the *step* (Special Test of English Proficiency), a 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 another study of rater differences, Kondo-Brown (2002) investigated whether trained native Japanese-speaking (JNS) raters would rate certain types of students and certain criteria more severely or leniently than others in assessing Japanese L2 compositions for norm-referenced purposes such as placement. She had three JNS raters rate 234 essays written by US university students studying Japanese as a foreign language. MFRM showed that the raters were self-consistent, but that they were significantly different from each other in their rating severity. Each rater had a different bias pattern for categories but was self-consistent across the categories of vocabulary, content, and mechanics. Each rater also displayed unique severe or lenient bias patterns towards test-takers. In spite of the presence of small but significant differences in rater severity and self-consistency in rater bias patterns, Kondo-Brown (2002) found no systematic overall bias patterns among the three raters. However, the percentage of significant rater–candidate bias interaction was much higher for candidates of extreme high or low ability.

In his study of systematic rater bias patterns in a Japanese EFL setting, Schaefer (2008) employed 40 native English speakers to rate 40 essays written by Japanese English majors. Each rater rated all the forty essays, using a six-point analytic rating scale consisting of five categories (Content, Organization, Style and Quality of Expression, Language Use, Mechanics, and Fluency). The results of his study using a many-facet Rasch model showed that "if Content and/or Organization were rated severely, then Language Use and/or Mechanics were rated leniently, and vice versa" (p. 465). He 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. Matsuno did not investigate a rater-item bias interaction.

Most recently, Winke, Gass, & Myford (2011) reported a study in which a group of 107 raters (mostly learners of Chinese, Korean, and Spanish) listened to a selection of 432 English speech samples that 72 test takers (native speakers of Chinese, Korean, and Spanish) produced. Their study examined whether a rater's prior study of a second language (L2) that matches the first language (L1) of a test taker influences the rater's rating of that test taker. MFRM revealed significant rater-writer bias interaction. They found that "when raters know, to varying degrees, the L1 of the test takers and can discern the L1 of the test takers though the test takers' ethnic accents, that knowledge may influence their ratings" (p. 53). Another interesting finding was that "raters were more lenient toward test takers

who had Spanish as their native language and more harsh toward test takers who had Korean or Chinese as their native language" (p. 53).

In investigating the phenomenon of rater subjectivity and inconsistency in second language 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.

There still do not seem to be many studies, however, which have investigated the possibility of systematic patterns in rater variation and no study has been done on systematic patterns in rater type variation. Du, Wright, & Brown's (1996) large-scale study (1734 essays and 89 raters) discovered systematic rater bias patterns, but individual essays were rated by only two raters each. Similarly, Engelhard (1992) examined 1000 essays with 82 raters, but with each essay rated by two raters. Kondo-Brown (2002) found that rater bias patterns were a matter of individual differences, but utilized only three raters. Wigglesworth (1993), with a larger number of raters (nine raters rating 36 tapes each for 83 candidates), also reported no common patterns, but a closer look at the 'assessment maps' for each rater reveals that there may be some discernable patterns. Some raters were harsh on grammar and lenient on fluency items, while other raters displayed the opposite pattern. Schaefer (2008) also found systematic variation in raters using rather a large number of native English speakers as raters. The only single study (Matsuno, 2009) which investigated rater variation in self-assessment, peer-assessment, and teacher assessment 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.

2 The Present Study

In the present study, MFRM was employed to investigate differential rater severity/leniency. First, we wanted to know whether our rating scale functioned reliably with rater type. We were also interested in how three rater types, namely self-assessor, peer-assessor, and teacher assessor, interact with assessment criteria or items of the rating scale. Closely related to this, we were also interested in the severity/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.

3 Research questions

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

1. How reliably does the rating scale function with three types of rater?
2. How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to items?
3. How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to students?
4. Are there any systematic bias patterns among self-assessors, peer-assessors, and teacher assessors?
5. Could student raters be used as an alternative for teachers for the purposes of essay rating?

III METHODOLOGY

1 Participants

The participants in the present study consisted of 194 raters, who were subdivided into student raters and teacher raters. Student raters were 188 undergraduate Iranian English majors enrolled in Advanced Writing classes in two state-run universities in Iran, comprising three fields of study: English Literature, Translation Studies, and English Language Teaching. The student raters were labeled either self-assessors or peer-assessors. Teacher assessors were six Iranian teachers of English. Following Dörnyei (2007), participants were selected according to the convenience sampling, which emphasizes the ease of accessibility of study participants.

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 fifty-seven (30.3%) were male. Eighty-six (45.7%) were native Farsi-speakers, 68 (36.2 %) were native-Turkish speakers, 4 (2.1%) were native-Kurdish speakers and another 4 (2.1%) were grouped as "Other". Ninety-five (50.5%) were sophomores, 29 (15.4%) were juniors, and 64 (34.0%) were seniors. Only 3 (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 years and most of them (61.7%) had studied the English language in language institutes before entering the university.

Teacher assessors were all male. They came from two language backgrounds: 4 teacher assessors were native-Farsi speakers, and the other 2 were native-Turkish speakers. They ranged in age from 23 to 36. 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, 1 of them with a private university, and 2 of them were classified as "Other". All of them had a degree in English: 3 of them were PhD students in ELT, 2had MAs in ELT, and 1had a BA in English literature.

2 The rating scale

Generally speaking, there are three types of rating scales in language testing: primary trait, analytic, and holistic (Weigle, 2002). Primary trait rating scales, as Weigle rightly puts it, "[have] not been widely used, and little information exists on how primary trait scoring might be applied in second language testing" (p. 110). Holistic rating scales do not represent writing well and are inadequate due to the provision of an overall, single score (Hamp-Lyons, 1991, 2003). For the purposes of the present study, we chose an analytic rating scale. The scale we developed for the present study is based on Jacobs, Zinkgraf, Wormuth, Hartfiel, &Hughey's (1981) ESL Composition Profile, but differs from it in many aspects (See appendix).

To develop our rating scale, we also consulted 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 for how to write the items for the scale: Composing With Confidence: Writing Effective Paragraphs and Essays (Meyers, 2006), Refining composition skills: grammar and rhetoric (Smalley, Ruetten, & Kozyreve (2000), and The Practical Writer with Readings (Bailey & Powel1, 2008). The scale contains fifteen 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).

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

As for the number of levels or bands, there is no consensus in the writing literature. Bond & Fox (2007) state that points from three to nine have been used and reported. For this study, we created a six-point scoring scale for each item. A six-point rating scale was chosen because these are "the most common number of scale steps in college writing tests, and a larger 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).

3 Data collection

One hundred and eighty eight five-paragraph essays were collected over a span of a year and a half from 188 students enrolled in Advanced Writing courses in two state-run prestigious universities in two different cities in Iran. The students in advanced writing classes are taught punctuation, expression, features of a well-written paragraph and principles of a one-paragraph and five-paragraph essay. This is the overall format of such a course as set by the Ministry of Sciences, Research, and Technology in Iran, and all instructors must follow this syllabus. The students were taught these principles of writing for eight weekly meetings. Immediately after the eight weekly meetings, they were told by their respective teachers they would have to sit the midterm exam the following week.

At the exam they were given 90 minutes to write a five-paragraph essay ranging in length from five hundred to seven hundred words on the following topic: ***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***. This topic was chosen from a list of TOEFL TWE topics. All the students were given one 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 one-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. During the rating process, the researcher monitored their ratings 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, new guidelines, and their own essays to rate. They knew that they were rating their own essays, and the researcher advised them to rate as accurately as possible. After the first rating, they were given their classmates' essays, with names removed, to rate as peer-assessment. The same procedure for self-rating was repeated for peer-assessment. The entire training and rating session took about two hours.

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

IV RESULTS

The present study employs a fully crossed design in which all raters rate all the essays, although this is not a requirement for MFRM. The only requirement is that there be sufficient overlap for data so that connectivity can be achieved (Linacre, 1989/1994, 2004). The data was analyzed with *Facets 3.68.0* a software program for MFRM (Linacre, 2011).Three facets were specified for this study: students, rater type, and items. The mathematical model for the facets is given below:

$\log(P_{nirk}/P_{nir}(k-1)) = B_n - D_i - T_r - F_k$

where:

P_{nirk} = the probability of student n being rated k on item i by rater type r,

$P_{nir}(k-1)$ = the probability of student n being rated k-1 on item i by rater type r,

B_n = the proficiency of student n,

D_i = the difficulty of item i,

T_r = the severity of rater type r, and

F_k = the difficulty of scale category k, relative to scale category k-1.

1 INITIAL ANALYSIS

Before answering the research questions, we did a preliminary Facets run to test for data-model fit. The results of the analysis showed that Students 94, 101, and 160, Raters 22, 24, 27, 48, 74, 76, 95, 145, and 176, and Item 7(logical sequencing) were misfits. The common practice in the literature (See McNamara, 1996) is to delete the misfitting elements. However, as the purpose of this study is to examine rater effects, and not to refine a test instrument, this approach was regarded as inappropriate, as we might end up throwing out the baby with the bath water. Unexpected ratings may reveal valuable insights into rater behavior, and so a different approach was adopted, which may be called a "lazer strategy" rather than a "scalpel strategy" (Myford, personal communication).

First we identified and deleted individual cases of highly unexpected ratings. We then reran the analysis and this time found no misfitting elements. According to Linacre (2011), satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are ≥ 2, and about 1% or less of (absolute) standardized residuals are ≥ 3. In our data, there were a total of 19,699 valid responses, that is, responses used for estimation of model parameters. Of these, 697 responses were associated with (absolute) standardized residuals ≥ 2, and 45 responses were associated with (absolute) standardized residuals ≥ 3, so the number of unexpected responses is much smaller than Linacre considers, indicating satisfactory model fit.

To answer the first question (How reliably does the rating scale function with three types of rater?), we present the following pieces of information. First, Figure 1 shows the vertical rulers, which compares all the facets on a common measurement scale.

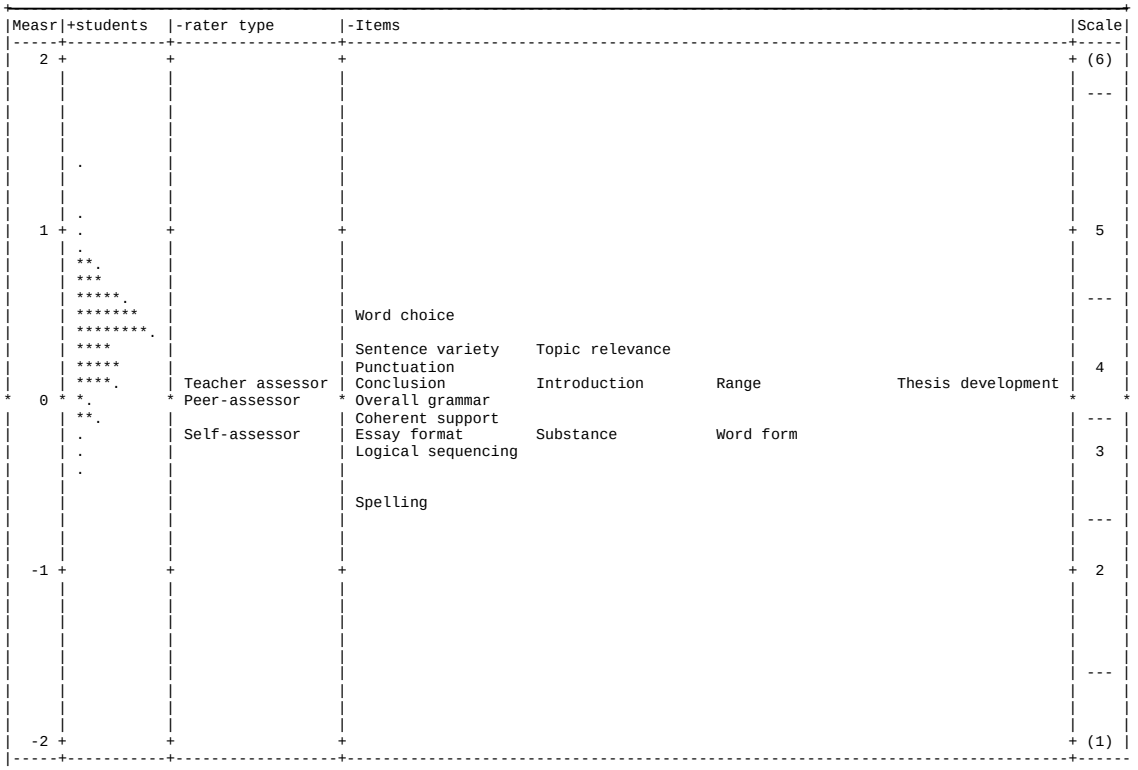

Figure 1. Variable map from the Facets analysis of the data

The first column in the map displays the logit scale, which ranges from 2 to -2 logits. The second column displays the student proficiency on the essay, with higher logits showing higher ability levels. Each star represents four students and each dot represents fewer than four students. The majority of students are located above the mean (which reflects positively on the instruction and test preparation they received. The third column displays the rater type. Although clustered around the mean, teacher assessor is the most severe rater type and self-assessor is the most lenient. The distribution of rater type is much narrower than that of students, which implies that we do not have to adjust student scores for individual differences of rater type. The fourth column displays the items. Items higher on the scale are more difficult and items lower on the scale are easier. Word choice is the most difficult item as scored by rater type across students. This finding is in contrast with previous studies such as McNamara (1996), Schaefer (2008), and Matsuno (2009), in which grammar was the most difficult item. One possible reason for this could be the setting in which the respective studies were done. McNamara's study was done in an ESL setting; those of Schaefer and Matsuno were done in EFL settings. Our study was also done in an EFL setting, Iran, but cultural differences between the two EFL settings, Japan and Iran, may be a mediating factor. Furthermore, the raters in McNamara and Schaefer's studies were native English speakers, but in our study rater type was composed of nonnative Farsi speaking students or teachers.

Spelling is the least difficult item as scored by rater type. This is in keeping with Matsuno (2009), who also found that spelling was the easiest item, and Hamp-Lyons (1991) notes that mechanics are considered to be superficial features of a piece of writing and might be easily ignored or unheeded. The fifth column displays the points or levels on the scale from one (very poor) to six (excellent).

The category statistics (Table 1) and the probability curves (Figure 2) also provide the answer to the first research question, indicating that the rating scale functioned reliably and validly with rater type.

Table 1 Category statistics for the rating scale

Data			Quality Control			Step Calibrations		
Category score	**Counts used**	**Cum. %**	**%**	**Average measure**	**Exp. measure**	**Outfit MnSq**	**S.E.**	**Most probable from**
1	875	4%	4%	-.03	-.08	1.0		
2	1877	10%	14%	.02	.05	.9	.04	-.78
3	3491	18%	32%	.17	.18	1.0	.02	-.51
4	5132	26%	58%	.31	.31	1.0	.02	-.14
5	5317	27%	85%	.44	.44	1.0	.02	.34
6	3007	15%	100%	.60	.59	1.0	.02	1.76

According to Linacre (2004), in order for a rating scale to function effectively, there should be at least ten observations in each category, average measures should advance monotonically with counts, outfit-mean squares should be less than two, and step difficulty or step calibration should advance by 1.4, but less than 5 logits. Table 1 shows that the rating scale meets all these guidelines: there are more than ten observations at each point, average measures advance monotonically, outfit mean squares are almost perfect (1.0), and the categories are the most probable ones, showing that the steps are appropriately ordered and function well (the guideline that step calibrations should advance by 1.4 is only true when the categories are dichotomous; otherwise, it can be ignored).

Figure 2 shows the student probability curves and is a graphic illustration of table 1. The probability curves show the threshold at which students are likely to be scored at the next highest level. These should resemble a range of hills (Linacre, 2004). That is, as ability level increases on the logit scale, the probability increases of achieving the next highest score ranking.

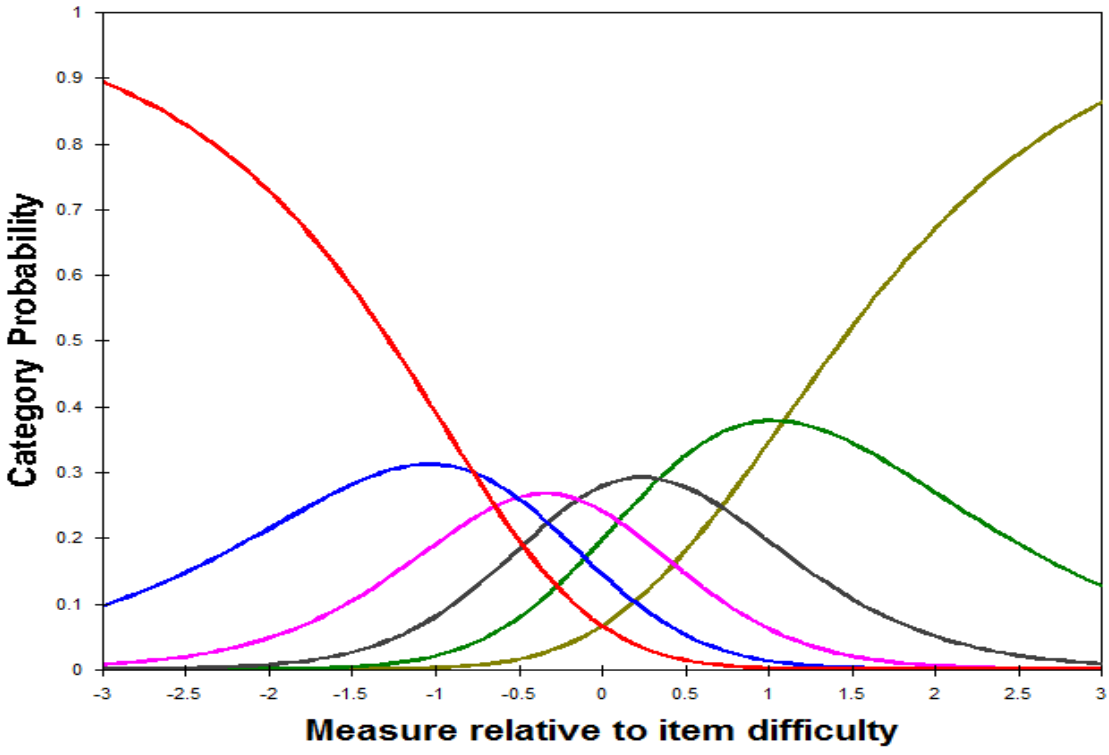

Figure 2 probability curves for students

Table 2 shows the calibration report for rater type. The first column of Table 2 shows the rater type by name. The second column shows the sum of the rater's ratings. The third column shows the total number of ratings each rater type gave. The fourth column shows the average of that rater type's ratings over all the examinees the rater type rated (Total Score/Total Count). The fifth column shows The Fair-M Average, which is an adjusted average for the rater type that takes into account two important factors: (1) the levels of proficiency of the particular students that the

rater type rated, and (2) the levels of severity of the other rater types who rated each of those particular students. In effect, the fair average corrects the observed average for the deviation of the students in each rater type's sample from the overall student mean across all raters. The sixth column shows the measure for the rater type, with the most lenient rater type (-16) at the top and the most severe rater type (12) at the bottom of the table. The seventh column shows the standard errors, which are very low, indicating a high precision of measurement. The last two columns show infit and outfit mean square for the rater type. Infit is an information-weighted sum, based on the sum of the squared standard deviations of the variance in a Rasch observation. It is inlier-sensitive—dominated by unexpected inlying patterns among informative, on-target observations. Outfit is based on the conventional sum of squared standardized residuals and is outlier-sensitive—dominated by unexpected outlying, off-target, low information low information responses (Myford & Wolfe, 2004).

Table 2 Calibration Report for Rater Type

Rater Type	Total Score	Total Count	Observd Average	Fair-Mean Average	Logit	Error	Infit	Outfit
Self-assessor	9543	2147	4.4	4.48	-.16	.02	.94	.95
Peer-assessor	8701	2089	4.2	4.18	.04	.02	.91	.91
Teacher assessor	62013	15463	4.0	4.06	.12	.01	1.02	1.5

Notes: Separation: 9.33, Strata: 12.78, Reliability: .99
Fixed (all same) chi-square: 205.4 d.f.: 2 significance: p< .00.

In order to examine the individual rater type fit statistics, we followed the suggested range for rating scale data of 0.5 to 1.5 in Wright, Linacre, Gustafson, & Martin-Lof (1994) (raters below this range overfit, or are too consistent; raters above this range underfit, or are excessively inconsistent). According to this range, none of the rater type showed misfit.

2 Rater type by items bias analysis

To answer the second research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency from each other in relation to items?), a bias analysis between rater type and items was specified in Facets. When we ran the analysis, there were 45 bias terms in all. Table 3 shows cases of significant rater type by items bias analysis.

Table 3 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	.6
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	1.0
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	1.4
Teacher	.12	10	-.22	4522	4428.8	.09	-.07	.03	-2.49	.9	.9

Fixed (all = 0) chi-square: 414.6 d.f.: 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, showing good precision of measurement. The mean square fit statistics are good, not showing any cases of misfit. Out of 45 bias terms, only 19 were statistically significant—with t-scores either greater than +2 or smaller than -2. 11 significant interactions are positive (showing severity), and 8 significant interactions are negative (showing leniency). The rater type showed significant bias toward only 10 out of 15 items. As is shown in table 3, rater type is biased toward items 2,3,4,6,7,8,10,13 and 15. 9 significant interactions are biased toward the self-assessor, 5 toward teacher assessor, and 5 others toward peer assessor. Self –assessor showed consistently severe bias most of the times (5 vs 4), peer-assessor and teacher assessor also showed consistently severe bias most of the times (3 vs 2). In other words, both peer-assessor and teacher assessor are equally consistently positively biased. All three rater types are severely biased toward items, but the number of times that self assessor shows cases of positive bias is greater than those of peer-assessor and teacher

assessor, so it could be argued that self –assessor is more severely biased than either peer-assessor or teacher assessor, although the sample size of bias significant bias interactions is too small. Item 3 (Topic relevance) showed the strongest rater type bias toward severity—bias toward teacher assessor (5.72) and the same item showed the weakest rater type bias toward leniency—toward self-assessor(-8.63).

3 Rater type by students bias analysis

To answer the third research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency from each other in relation to students?), another bias/interaction analysis similar to rater type-items bias analysis was specified for rater type and students in Facets. Table 4 shows cases of rater type-by-students bias analysis. The standard errors (SEs) are low, showing good precision of measurement. Compared to the rater-items bias analysis, SEs are much greater especially for student raters, but they are still low. 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 4 Rater type-items bias/interaction analysis

Rater type	Logit	Students	logit	Obs score	Exp score	Obs-Exp score	Bias size	Model S.E.	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	.9
Peer	.05	16	.21	57	45.7	.94	-.68	.27	-2.48	.7	.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
Self	-.17	29	.55	72	60.7	.87	-1.12	.41	-2.71	.8	.7
Peer	.05	29	.55	77	65.5	.76	-.70	.29	-2.46	1.1	1.1
Self	-.17	31	-.21	59	49.1	.71	-.45	.22	-2.02	.7	.8
Self	-.17	32	.28	74	64.2	.65	-.54	.26	-2.09	.8	.9
Self	-.17	33	.00	78	57.9	1.34	-1.13	.30	-3.80	.7	.7
Peer	.05	34	.57	77	65.9	.74	-.69	.29	-2.40	.8	.7
Self	-.17	36	.82	60	68.9	-.64	.53	.22	2.36	.7	.7
Self	-.17	37	-.07	67	56.3	.72	-.48	.23	-2.13	.3	.3
Self	-.17	38	.21	77	62.7	.96	-.84	.29	-2.93	.5	.6
Peer	.05	51	.20	26	53.6	-1.97	1.36	.28	4.95	.9	.7
Peer	.05	52	.25	43	58.8	-1.05	.64	.20	3.12	.8	.8
Peer	.05	53	.40	73	62.2	.72	-.57	.25	-2.23	1.0	.9
Self	-.17	55	.37	55	66.0	-.74	.48	.20	2.41	.6	.6
Peer	.05	55	.37	49	61.5	-.83	.51	.20	2.57	.2	.2
Peer	.05	56	.02	31	48.6	-1.26	.81	.24	3.40	.3	.3
Peer	.05	58	1.02	87	73.8	.88	-1.75	.58	-3.01	.8	.9
Teacher	.12	58	1.02	270	286.5	-.28	.25	.12	2.10	1.6	1.6
Peer	.05	62	.38	50	61.9	-.79	.49	.20	2.45	1.5	1.5
Self	-.17	64	.82	64	74.0	-.66	.55	.22	2.55	.7	.7
Self	-.17	66	.53	54	69.3	-1.02	.69	.20	3.46	.7	.7
Self	-.17	70	.09	75	59.9	1.00	-.80	.27	-3.00	1.0	1.0
Self	-.17	72	.94	55	75.7	-1.38	1.06	.20	5.29	.4	.4
Teacher	.12	72	.94	381	357.2	.32	-.32	.12	-2.59	1.4	1.3
Self	-.17	73	.56	76	64.8	.80	-.94	.36	-2.60	.8	.8
Peer	.05	77	.25	76	58.7	1.15	-.93	.28	-3.37	.4	.4
Self	-.17	79	.08	38	59.8	-1.46	.90	.21	4.20	.3	.3
Teacher	.12	79	.08	347	317.0	.33	-.20	.08	-2.44	1.2	1.2
Self	-.17	80	.06	44	59.3	-1.02	.62	.20	3.05	.6	.6
Peer	.05	80	.06	43	54.2	-.74	.45	.20	2.19	.3	.3
Teacher	.12	80	.06	337	310.5	.30	-.18	.08	-2.16	1.1	1.2
Peer	.05	81	.50	47	60.5	-.97	.61	.21	2.96	.3	.3
Peer	.05	82	.36	37	61.5	-1.63	1.01	.22	4.69	.5	.5
Peer	.05	83	.37	45	61.7	-1.11	.68	.20	3.37	.3	.3
Self	-.17	84	.52	53	69.0	-1.07	.72	.20	3.59	.7	.7
Peer	.05	84	.52	77	61.4	1.12	-1.27	.38	-3.30	1.2	1.2
Peer	.05	85	.59	81	66.3	.98	-1.05	.34	-3.10	.7	.7
Self	-.17	86	.34	46	65.6	-1.31	.82	.20	4.09	.3	.3
Self	-.17	87	.38	50	66.3	-1.09	.70	.20	3.50	.4	.4
Peer	.05	88	-.12	33	49.5	-1.10	.73	.23	3.13	.6	.6

Self	-.17	89	.56	57	69.8	-.86	.60	.20	2.96	.5	.4
Peer	.05	92	.40	41	62.2	-1.42	.87	.21	4.21	.6	.7
Self	-.17	97	.10	71	60.3	.71	-.53	.24	-2.17	1.2	1.3
Peer	.05	101	.06	40	54.1	-.94	.57	.21	2.74	1.4	1.3
Peer	.05	104	.17	79	56.8	1.48	-1.27	.31	-4.10	1.3	1.3
Teacher	.12	104	.17	249	274.5	-.34	.20	.09	2.28	.9	.9
Peer	.05	109	.40	49	62.3	-.89	.55	.20	2.75	.7	.7
Peer	.05	113	.32	47	60.5	-.90	.55	.20	2.74	.5	.5
Peer	.05	120	.67	77	67.9	.60	-.58	.29	-2.04	.9	.9
Self	-.17	125	.64	59	71.1	-.81	.59	.21	2.88	.8	.8
Self	-.17	128	.67	76	66.6	.67	-.83	.36	-2.30	.6	.5
Peer	.05	128	.67	50	63.5	-.96	.64	.21	3.10	.5	.5
Self	-.17	134	.47	57	68.0	-.74	.50	.20	2.48	.3	.3
Peer	.05	135	.81	90	70.5	1.30	-4.00	1.60	-2.50	.0	.0
Self	-.17	139	.59	83	70.3	.85	-1.10	.38	-2.87	.6	.7
Self	-.17	147	.48	78	68.4	.64	-.65	.30	-2.17	1.0	1.0
Self	-.17	150	.36	76	65.9	.67	-.60	.28	-2.18	.9	1.0
Peer	.05	152	.66	50	67.7	-1.18	.77	.20	3.85	.3	.3
Peer	.05	155	-.08	27	50.7	-1.58	1.15	.27	4.19	.2	.3
Teacher	.12	155	-.08	264	240.3	.32	-.19	.09	-2.12	1.1	1.1
Self	-.17	158	.24	85	63.4	1.44	-1.80	.45	-3.97	.6	.6
Peer	.05	158	.24	78	58.6	1.29	-1.10	.30	-3.70	.4	.4
Teacher	.12	158	.24	235	276.0	-.56	.34	.09	3.73	1.3	1.3
Self	-.17	161	-.11	38	55.2	-1.14	.70	.21	3.30	1.0	1.0
Self	-.17	162	.46	78	67.8	.68	-.67	.30	-2.27	.5	.5
Peer	.05	165	-.06	67	51.2	1.05	-.68	.23	-3.04	1.1	1.0
Self	-.17	166	.30	53	64.7	-.78	.50	.20	2.51	.7	.8
Peer	.05	168	.56	82	65.6	1.09	-1.21	.36	-3.36	.7	.7
Self	-.17	169	.41	85	66.9	1.21	-1.63	.45	-3.60	.7	.7
Peer	.05	172	.27	46	59.2	-.88	.53	.20	2.65	.9	.9
Self	-.17	173	.52	79	69.0	.67	-.70	.31	-2.27	.7	.9
Self	-.17	176	.57	85	70.0	1.00	-1.46	.45	-3.24	.9	.9
Self	-.17	177	.00	37	57.8	-1.39	.86	.22	3.97	.7	.7
Peer	.05	177	.00	77	49.6	1.96	-1.82	.38	-4.73	.8	.8
Peer	.05	181	.18	40	57.1	-1.14	.69	.21	3.32	.3	.2
Peer	.05	182	.77	57	69.7	-.84	.59	.20	2.91	.7	.7
Self	-.17	185	.38	85	66.4	1.24	-1.65	.45	-3.66	1.2	1.4
Self	-.17	187	.25	78	63.6	.96	-.88	.30	-2.97	1.2	1.1
Peer	.05	188	.33	50	60.8	-.72	.44	.20	2.21	.3	.3

The mean square fit statistics are good, but, unlike rater type-items bias analysis, there are misfits in rater type-students bias analysis. As is displayed in Table 4, out of 472 bias terms, only 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 forty-five are positive (showing severity). Rater type showed significant bias toward all the students, and in three cases all the rater types showed bias toward the same student—students 8, 80, and 158; but only 65 out of 188 students had significant bias interactions toward rater type. A similar pattern was found in Schaefer (2008).

4 Patterns in rater type

Tables 5 and table 6 answer the fourth research question (Are there any systematic bias patterns among self-assessors, peer-assessors, and teacher assessors?). In order to explore possible systematic patterns in rater type-items and rater type-students bias interactions, we arranged the significant bias interactions to show the frequency patterns in tables 5 and 6. Table 5 shows the rater type-items relationship. The first two columns show the items and item numbers, the third column shows the logit for the items, the next three columns show the rater type, the seventh column shows the total number of patterns for rater type and finally the last two items show leniency/severity of rater type. As shown in table 5, one interesting result is that student raters (self and peer) show the opposite pattern of severity/lenience as 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.

A closer inspection of Table 5 shows that out of 19 misfitting rater types, seven belong to self-assessors, 11 to peer-assessors and one to 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 or abilities range from -0.07 to .66 for student raters, from -0.07 to .56 for self-assessors, and from 0.02 to .66 for peer-assessors. 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, Linacre et al (1994) claim we should first treat underfits because they force elements to remain below 1. They also claim that the overfitting elements should remain in the analysis because

although they do not provide something 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.

Table 6 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, with a spread of 1.80 logits. Across the top of the table is the student logit range, form the lowest ability, -.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 5 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

Table 6 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

As can be seen in table 6, students do not show enough spread—1.80 logits, and the majority of students (163) cluster above mean between 0.00 and 1.00. There are only 18 students falling below mean at the lower end of the logit scale and there are only seven falling above 1.00 at the upper end of the logit scale. This can be justified due to the very fact that students were well taught for about eight weekly meetings the principles of essay writing, which shows that 163 out of 188 students were well prepared, showing good ability.

The majority of bias interactions, 81 out of 91, fall above the mean, between 0.00 and 1.00, 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. Another noteworthy point is that 46 bias interactions are negative (showing leniency) and 45 are positive (showing severity). This shows that rater type is consistently leniently biased toward the students. The third point concerning the table is that rater type shows bias toward higher ability students (between 0.00 and 1.00), with slightly more cases of lenient bias (41 vs. 40). 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 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 3) 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. As can be seen, some interesting patterns emerge for both lowest and highest ability groups, but again, the small number of cases makes generalization difficult.

When patterns are compared, 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 (2009:75) who concluded that "self-assessment was somewhat idiosyncratic and therefore of limited utility as a part of formal assessment."

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 severely biased

toward students, which is in line with Handrahan & Issacs (2001) who also found that peers could be very critical, but teacher assessors show severe and lenient bias in equal measure.

V Discussions and conclusions

The present study set out to investigate whether there could be any interactions between three types of assessors and students or items and whether these interactions could be systematically patterned. The present study further intended to argue for a place for student raters in essay rating in higher education. The findings of the study showed some recurring patterns. Two types of bias were found in this study: rater type by items and rater type by students. These are explained in detail below.

Student raters (self and peer) show the opposite pattern of severity/lenience as the teachers toward items. 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. There is only one reason as it applies to this study. 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 studies (Lumley, 2005). The monitoring made student assessors rate similar to each other and have similar patterns to each other. This has been proved to result in consistency (See, for instance, Knoch, Read, and Von Randow 2007; Saito, 2008).

Both self-assessors and teacher assessors show the opposite pattern of severity/leniency as peer-assessors toward the extreme ends of student ability groups. This is in line with Schaefer (2008) and is further discussed below. Where both self-assessors and teachers assessors are lenient, peer-assessors are severe and vice versa.

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 are so critical of their peers and is line with many previous studies including Handrahan & Issacs (2001).

Spelling is the easiest item as scored by rater type. This finding is consonant with Matsun0 (2009), and it is because the superficial features are usually not given in-depth thought (Hamp-Lyonz, 2003). 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 be argued in relation to the setting in which the respective studies were done. It seems that different studies at different settings using different raters produce different results concerning the most difficult item and could be attributed to the perceptions and 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 setting, the former using rather inexperienced native English speaking raters while the latter using student raters. Another possible interpretation, as it relates to the present discussion and as has been confirmed in previous studies (Saito and Fujita, 2009), might be because raters, especially student raters, are generous in their rating of some items or are unable to differentiate items, hence resulting in inflated marking. In passing, it should be noted that when separate analyses were run, word choice was the most difficult item as scored by rater type, but spelling was the easiest item shared only by peer-assessor and teacher assessor, but not by self-assessor. This further confirms the reason given above.

The present study is inconclusive in answering the question of whether self- or peer-assessment could be an alternative to teacher assessment in awarding grades on essay writing. This finding deserves further elaboration, as it is an answer to the last research question too (Could student raters be used as an alternative for teachers for the purposes of essay rating?). There are some inconsistencies. When we look at table 5, both self assessors and peer assessors rate very similar 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 an opposite pattern. Table 6, however, reveals a different pattern. Here self assessors and peer assessors rate almost differently and self assessors rate similar to teacher assessors. Self assessors tend to overrate themselves and this has been shown in previous research in which low ability students tend to overrate themselves and high ability students tend to underrate themselves (Blanche & Merino, 1989; Boud & Faichikov, 1989; Falchikov & Boud, 1989), which could be attributed to experience (Ross, 1998). Student raters are, however, consistent when it comes to assessment criteria, which is in keeping with (Falchikov & Goldfinch, 2000)who conclude that when the criteria are explicitly stated and well understood, they lead to more accurate and consistent marking of 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, p. 31) 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." Another plausible interpretation for the inconsistency of the results could be due to the lack of research using MFRM in this area, as Matsuno (2009) rightly puts it, "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 is another reason for the inconsistency. Ross (1998) is an exception in self-assessment which mainly focuses on self assessment and four language skills or language prfoiciency, not necessarily writing product.

The results of this study could lead to two implications: one for research in rater training and the second one for pedagogy in concurrent validity of ratings in L2 writing classrooms. One hour training coupled with monitoring in the present study led to more consistency on 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 for the similarity between self, peer, and teacher rating, it was shown in the study that self and teacher ratings were similar to each other, which provides partial evidence for concurrent validity of the self and teacher ratings.

Although the findings of this study could be encouraging and promising, a few caveats and limitations are in order. This study was done purely quantitatively. Lack of a qualitative component failed to provide us with cogent and justifiable reasons why the 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 that of other studies in which self and peer assessments were involved. Still, caution needs to be exercised as for the generalizability of the results. The third limitation is the small number of essays both self-assessors and peer-assessors rated (one essay each). The last limitation concerns the language proficiency of students. Although proficiency of students was irrelevant in the present study, heterogeneity of students could have affected the essays and ratings. 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. Future studies should strive to answer this important question.

Acknowledgments

This study is based on the final report of a research project toward the PhD thesis of the second author and was financially supported by the Research Office of the University of Tabriz. The grant was given to the first two authors.

REFERENCES

Bailey, E. P. and P.A. Powel1, 2008. The practical wrier with readings. United States of America, Thomson/Wadsworth.

Blanche, P. and B. J. Merino, 1989. Self assessment of foreign language skills: implications for teachers and researchers. Language Learning, 39(3): 313-340.

Bond, T.G. and C. M. Fox, 2007. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah NJ, Lawrence Erlbaum.

Boud, D. and N. Falchikov, 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. Higher Education, 18 (5): 529-549.

Dörnyei, Z. 2007. Research methods in applied linguistics : Quantitative, qualitative, and mixed methodologies. Oxford, Oxford University Press.

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

Eckes, T, 2009. Many-facet Rasch measurement. Retrieved June 1, 2011, from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf

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

Engelhard, G. Jr., and G. E. Stone, 1998. Evaluating the quality of ratings obtained from standard-setting judges. Educational and Psychological Measurement, 58(2): 179–196.

Falchikov, N. and D. Boud, 1989. Student self assessment in higher education. Review of Educational Research, 59 (4): 395-430.

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

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

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

Hamp-Lyons, L, 1995. Rating nonnative writing: The trouble with holistic scoring. TESOL Quarterly, 29(4): 759-762.

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

Handrahan, S. and G. Issacs, 2001. Assessing self- and peer-assessment: The students' views. Higher Education research and Development, 20(1): 53-70.

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

Knoch, U., J. Read, J., and T. von Randow, 2007. Re-training writing raters online: How does it compare with face-to-face training? Assessing Writing 12(1): 26–43.

Kondo-Brown, K, 2002. A FACETS analysis of rater bias in measuring Japanese L2 writing performance. Language Testing, 19(1): 3–31.

Linacre, J. M, 2011. FACETS (Version 3.68.0) [Computer Software]. Chicago, IL, MESA Press.

Linacre, M, 1989/1994. Many-facet Rasch measurement. Chicago, MESA press.

Linacre, M, 2004. Optimizing rating scale effectiveness. In: Introduction to Rasch model., (Eds E. V., Smith, Jr, and R. M. Smith.). pp. 258-278. Maple Grove, Minnesota, JAM press.

Lumley, T, 2002. Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3): 246–276.

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

Matsuno, S, 2009. Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. Language Testing, 26(1): 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. and E. W. Wolfe, 2004. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In: Introduction to Rasch measurement. (Eds E. V. Smith and R. M Smith.) pp. 460–517. Maple Grove, MI, JAM Press.

O'Neill, T. R. and M. E**.** Lunz, 1996. 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 1996, New York.

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

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

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

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

Schaefer, E, 2008. Rater bias patterns in an EFL writing assessment. Language Testing, 25(4): 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(3): 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.

Winke, P., S., Gass, and C. Myford, 2011. The relationship between raters' prior language study and the evaluation of foreign language speech samples (TOEFL iBT Research Report TOEFL iBT-160). Princeton, NJ, Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-30.pdf

Wright, B. D., J. M., Linacre, J. -E., Gustafson, and P. Martin-Lof, 1994. Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370. Retrieved June 1, 2011, from http://rasch.org/rmt /rmt83b.htm

APPENDIX: Essay Rating Sheet

Essay number:						
Assessor'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 ☐