Iranian Journal of Applied Linguistics (IJAL), Vol. 16, No. 1, March 2013, 145-175

Rater Bias in Assessing Iranian EFL Learners' Writing Performance

Mahnaz Saeidi [a]
Associate Professor of Applied Linguistics, Tabriz Branch, Islamic Azad University, Tabriz, Iran

Mandana Yousefi [b]
PhD Graduate of TEFL, Tabriz Branch, Islamic Azad University, Tabriz, Iran

Purya Baghayei [c]
Assistant Professor of Applied Linguistics, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Received 23 September 2012; revised 13 January 2013; accepted 7 February 2013

Abstract

Evidence suggests that variability in the ratings of students' essays results not only from their differences in their writing ability, but also from certain extraneous sources. In other words, the outcome of the rating of essays can be biased by factors which relate to the rater, task, and situation, or an interaction of all or any of these factors which make the inferences and decisions made about students' writing ability undependable. The purpose of this study, therefore, was to examine the issue of variability in rater judgments as a source of measurement error; this was done in relation to EFL learners' essay writing assessment. Thirty two Iranian sophomore students majoring in English language participated in this study. The learners' narrative essays were rated by six different raters and the results were

[a] *Email address:* m_saeidi@iaut.ac.ir; mnsaeidi@yahoo.ca
Corresponding address: Department of English, Faculty of Persian Literature and Foreign Languages, Tabriz Branch, Islamic Azad University, Tabriz, Iran

[b] *Email address:* m_yusefi@yahoo.com
[c] *Email address:* puryabaghaei@gmail.com

analyzed using many-facet Rasch measurement as implemented in the computer program FACETS. The findings suggest that there are significant differences among raters concerning their harshness as well as several cases of bias due to the rater-examinee interaction. This study provides a valuable understanding of how effective and reliable rating can be realized, and how the fairness and accuracy of subjective performance can be assessed.

Keywords: Rater bias; Writing ability; Many-Facet Rasch Measurement; Inter-rater reliability

Introduction

It is common practice to describe learners' achievements on the basis of test scores. Studies often report differences in test scores between subgroups of an entire population. Of course, differences found among the learners may be caused by the fact that they differ in their command of the skills the test intends to measure. However, they may also be wholly or partially caused by the measuring procedure used. When assessing the writing of their students, the teachers expect to find different writing skills and abilities and to give out different scores. No matter what the method or the test for assessment is, the reliability of ratings is one of the major issues in assessing writing ability (McNamara, 1996). As we have always observed, there exists variance in the ratings of students' writing tasks due to differences in their writing abilities. However, the findings of different studies have shown that the learners' scores can be affected by factors not related to the ability being assessed and this introduces the concept of bias in assessment.

The research literature indicates that bias in general has attracted the interest of many researchers (some other studies include Congdon, 2006; Engelhard, 2002; Lumley & McNamara, 1995; Lunz, Stahl, & Wright, 1991; Lynch & McNamara, 1998; McNamara & Adams, 1991; Moon & Hughes, 2005; Nijveldt et al., 2009; O'Neill & Lunz, 1997), but most of these researchers studied bias resulting from sex, race, ethnic group, social status, or other factors that caused discrimination among different groups in society, and few have examined bias resulting from an interaction between rater and some facet concerned with the examinee or test (rater-examinee, rater-rating scale, or rater-task). At the same time few of the studies reviewed, with the exception of Kondo-Brown (2002), Schaefer (2008), and Eckes (2005, 2012) conducted their research in an EFL context. With regard to the fact that in an EFL context there is a limited exposure to English language outside the classroom and the learners' development of writing ability is to a large extent

dependent on the teacher's instruction and assessment in the class, and of course the fact that raters' inconsistencies in rating would result in unfair educational decisions which are not desirable for all the teachers, this study, therefore, has made an attempt to investigate the sources of raters' inconsistencies with regard to writing assessment in Iranian EFL context. The purposes of this study were three-fold: Firstly, to determine the degree of differences among the raters concerning the rating of learners' essays; secondly, to investigate the existence of rater bias due to rater-rating scale or rater-examinee facets; and finally, it was the researcher's concern to examine whether there was any difference among the raters with regard to the rating scale dimensions.

Background

Conceptual Definition of Bias

According to Sudweeks, Reeve, and Bradshaw (2005), a student's score on a given essay will be also influenced by several extraneous factors including:

> (a) The nature of the particular writing prompt or task posed, (b) the particular rater(s) who judged the student's essay, (c) situation-specific factors associated with the particular rating occasion, (d) the student's background and interest in the topic or problem presented, and (e) interactions among these different sources. (p. 240)

The variability resulted from these extraneous sources is considered to be measurement error and the test involving such a kind of variability is biased. It is obvious that the teacher does not aim to make the decision about his learners based on the scores not exactly showing the evaluation of the desired ability or some extraneous factors and this study is an attempt to investigate the sources of raters' inconsistencies with regard to writing assessment in Iranian EFL context. As Schaefer has noted:

> The idea of searching for unexpected interactions among rater judgments and test takers' performance or other facets in an analysis is central to bias analysis. It 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. (2008, p. 467)

Bias in assessment conveys "a skewed and unfair inclination toward one side (group, population) to the detriment of another" (McNamara & Roever, 2006, p. 82) and is directly related to fairness. Bias can be seen in traditional validity terms

as 'construct-irrelevant variance that distorts the test results and therefore makes conclusions based on scores less valid' (McNamara & Roever, 2006, p. 82). Accordingly, if learners of equal ability score differently on a test or item, there exists a construct-irrelevant variance which affects the learners' scores, causing the unidimensional test to become multidimensional. Thus, the test measures not only what it is intended to measure but something more, making the result an invalid source for interpretation. Biased tests harm all the educational and social institutions, since students might be admitted to a program or job for which they do not have the required ability and knowledge, while, on the other hand, qualified individuals might be rejected and deprived of their deserved positions and rights.

According to Van de Vijver and Tanzer (2004), bias occurs if score differences on the indicators of a particular construct do not correspond to differences in the underlying trait or ability. Van de Vijver and Poortinga (1997) distinguished three kinds of bias:

The first one is construct bias which occurs if the construct measured is not identical across cultural groups. Western intelligence tests provide a good example of this. In most general intelligence tests, there is an emphasis on reasoning, acquired knowledge, and memory, with social aspects of intelligence being less often emphasized. However, there is ample empirical evidence that these aspects may be more prominent in non-Western settings. Thus, the use of Western intelligence test for non-western subjects is an example of construct bias.

The second one is method bias which includes sample bias, instrument bias, and administration bias. Sample bias arises from incomparability of samples on aspects other than the target variable. For instance, intergroup differences in motivation can be a source of method bias caused by sample incomparability (subjects frequently exposed to psychological tests show less motivation than subjects for whom the instrument has high novelty). Instrument bias refers to problems deriving from instrument characteristics or response procedure. A well-known example is stimulus familiarity. Deregowski and Serpell (1971) asked Scottish and Zambian children to sort miniature models of animals and motor vehicles as well as photographs of these models. Although no cross-cultural differences were found for the actual models, the Scottish children obtained higher scores than the Zambian children when photographs were sorted. Administration bias arises when, for example, with these interviewees, there is insufficient knowledge of the testing

language, or when inappropriate modes of address or cultural norm violations are used by the interviewer; these factors make the collection of appropriate data impossible.

The final type is item bias which refers to distortions at item level. Biased items have a different psychological meaning across cultures. For participants from different cultural groups who are equal concerning whatever is measured, an unbiased item should be equally difficult and they should have equal mean scores across the cultural groups; different means on that item refers to item bias.

Empirical Studies on Bias

Regarding bias analysis, many studies have found unexpected interactions between rater judgments and other facets not related to test takers' performance. For example, Wigglesworth (1994) looked at rater-item, rater-task, and rater-test type interaction in the speaking test for potential immigrants to Australia; she found significant rater differences in the way candidates responded to different items. Some raters were consistent in their overall ratings, while others rated grammar, fluency and/or vocabulary either more harshly or more leniently. Raters could also be differentiated by their harshness or leniency towards different task types.

In Australia, McNamara (1996) found that trained raters were overwhelmingly influenced by candidates' grammatical accuracy in the Occupational English Test. While the grammatical accuracy was important according to Many-Facet Rasch Measurement (MFRM), this was regarded as of little importance by the raters; that is, there was a difference between what the raters thought they were doing, and what they actually did. McNamara concluded that "Rasch analysis is useful in revealing underlying patterns in ratings data which can be interpreted in ways that raise fundamental questions of test validity" (1996, p. 216).

Lumley (2002) used MFRM to analyze the writing component of the Special Test of English Proficiency (STEP) for immigrants to Australia and found significant differences between raters' severity toward rating grammar.

In another study, Kondo-Brown (2002) investigated trained native Japanese-speaking (JNS) raters' severity in assessing U.S. university students' Japanese L2 compositions. Three JNS raters rated 234 essays written by students studying Japanese as a foreign language. Using MFRM, she concluded that the raters were

significantly different from each other in their rating severity. Each rater had a different bias pattern for different dimensions but was self-consistent across the dimensions of vocabulary, content, and mechanics. Kondo-Brown found no systematic overall bias pattern among the three raters. However, the percentage of significant rater–candidate bias interaction was much higher for candidates of extreme high or low ability.

Eckes (2005) studied rater effects in the writing and speaking sections of the Test of German as a Foreign Language (Test DaF). Focusing on rater main effects as well as interactions between raters, examinees, rating criteria, and tasks, he found that raters (a) differed strongly in the severity with which they rated examinees; (b) were fairly consistent in their overall ratings; (c) were substantially less consistent in relation to rating criteria than examinees; and (d) as a group, were not subject to gender bias.

Schaefer (2008) employed MFRM to explore the rater bias patterns when they rate 40 essays written by female Japanese university students on a single topic. The results revealed several recurring bias patterns among rater subgroups. Regarding rater–category bias interactions, "twenty-four out of the 40 raters had significant bias interactions with categories, and there were 57 significant bias terms in all. Twenty-seven of the significant bias interactions were negative (showing leniency), and 30 were positive (showing severity)" (p. 480). In addition, if Content and/or Organization were rated severely, then Language Use and/or Mechanics were rated leniently, and vice versa. In rater–writer bias interactions, raters were either more severe or more lenient towards higher ability writers than lower ability writers. In sum, 329 significant rater-writer interactions were observed among which164 interactions tended towards unexpected severity, and 165 tended towards unexpected leniency.

Taking classical test theory and MFRM model as the theoretical basis, Haiyang (2010) investigated the reliability of an English test for non-English major graduates. The results showed that the candidates' scores of the objective test were not significantly correlated with their scores of the subjective tasks. The results of the MFRM analysis indicated that the raters' severity difference in their rating, the varying difficulty levels of the test tasks, and the bias interaction between some students and certain tasks caused the variance in the scores.

In order to examine the raters' severity/leniency regarding criteria, Eckes (2012) investigated the relation between rater cognition and rater behavior. Based on the ratings of 18 raters, criterion-related bias measures were estimated using MFRM which yielded four operational rater types. He concluded that "criteria perceived as highly important were more closely associated with severe ratings, and criteria perceived as less important were more closely associated with lenient ratings" (p. 270).

Many-facet Rasch Measurement

A term coined for technical analyses of test items and detecting biased test items is Differential Item Functioning (DIF). According to McNamara and Roever (2006):

> DIF identifies test items that function differently for two groups of test takers and is a necessary but not sufficient condition for bias because a test item that functions differently for two groups might do so because it advantages one group in a construct-irrelevant way, but there might also be legitimate reasons for differential functioning. (p.83)

DIF has not been employed for detecting bias in tests like essays which do not include different items. Instead, most of the researchers dealing with performance assessment (Lumley & McNamara, 1995; Lynch & McNamara, 1998; Schaefer, 2008; Sudweeks, Reeve, & Bradshaw, 2005; Weigle, 1998) have used MFRM which is an extension of the Rasch model (Rasch, 1980) developed by Linacre (1989).

Classical test theory (CTT) provides several ways of estimating reliability by distinguishing true scores from error scores. Sources of error scores might include random sampling error, internal inconsistencies among test items or tasks, and inconsistencies over time, across different forms of test or within and across raters. According to Haiyang (2010), CTT estimates of reliability have several limitations. Firstly, CTT estimates cannot provide information about the effects of multiple sources of error and how these differ. Secondly, CTT treats all errors to be random or unidimensional and do not distinguish systematic measurement error from random measurement error. Finally, CTT has a single estimate of standard error of measurement for all candidates. The early efforts at investigating bias, classical test theory indices and ANOVA approaches are no longer considered appropriate for

studying items, because "mean differences in performance are confounded with item difficulty" (Camilli & Shepard, 1994, p. 25).

Item Response Theory (IRT) includes a range of probabilistic models for describing the relationship between a test taker's ability level and the probability of his or her correct response to any individual item (Shultz & Whitney, 2005). Item response theory differs from classical test theory by modeling the interaction of the person and the individual items to a latent trait. By modeling responses in terms of their relations to a common underling trait, IRT models have an important feature that allows us to determine if people from two groups respond differently to the same item given that they have the same level of a trait (Bolt & Rounds, 2000, as cited in Einarsdóttir & Rounds, 2009). IRT rests on the premise that a test taker's performance on a given item is determined by two factors: The test taker's level of ability and the characteristics of the item. MFRM (Linacre, 1989) is an extension of one-parameter Rasch model (Rasch, 1980), which is a special case of IRT model, 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). It enables us to include multiple aspects, or facets, of the measurement procedure in the test results analysis. A facet of measurement is an aspect of the measurement procedure which the test developer believes may affect test scores and hence needs to be investigated as part of the test development (e.g. task or item difficulty, rater severity, rating condition, etc.). MFRM has been used by many researchers to investigate rater bias in a number of studies. It enables the 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, and thus, it can analyze sources of variation in test scores besides item difficulty or person ability. 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 (Engelhard, 1992).

Sudweeks, Reeve, and Bradshaw (2005) describe the original Rasch model as a model in which "the persons and test items are evaluated and placed on an equal-interval scale in terms of their differing abilities (persons) or difficulties (items). The results are sample-independent" (p. 243).

According to Lumley and McNamara (1995, as cited in Sudweeks, Reeve, & Bradshaw, 2005), MFRM which is implemented through the computer program FACETS, allows assessing the effects of different sources of systematic errors in the ratings such as,

> inconsistencies between raters, differences in ratings between rating occasions, and differences in the relative difficulty of various writing tasks (prompts). It provides information about how well the performance of each individual, rater, or task matches the expected values predicted from the model generated in the analysis. These *fit statistics* are known in *Rasch analysis* as *infit* and *outfit* mean square values. (p. 243)

The Present Study

This study aimed at investigating the degree of differences among the raters in terms of the rating of learners' essays, the existence of rater bias because of rater-rating scale or rater-examinee facets, and differences among the raters with regard to the rating scale dimensions.

Accordingly, the following research questions have been formulated:

1. To what degree do the raters differ from each other in their assessments of EFL learners' writing ability?
2. Does the interaction between raters and examinees cause bias in raters' assessment of EFL learners' writing ability?
3. Does the interaction between raters and rating scale cause bias in raters' assessment of EFL learners' writing ability?
4. Are there any systematic bias patterns due to rater-rating scale or rater-examinee facets among raters?
5. Does the raters' rating differ from each other regarding the rating scale category characteristics?

Method

Participants

Thirty two Iranian sophomore university students majoring in English Translation and English Literature at Islamic Azad University, Quchan Branch participated in this study. All the students chose Advanced Writing Course as the requirement of the third semester of their major. Since all the courses in the first three semesters of English Translation and English Literature are the same, the major was not

regarded as a likely facet in the analysis. This exploratory study had a non-probability sampling design. Both male and female students participated in the study and the students' age range was 20-27.

Instrumentation

A Test of essay writing was administered to give learners a chance to compose, under a forty five minute time constraint, a narrative essay about 'A happening in my childhood'. The topic and administration of the test was based on TWE (Test of Written English in TOEFL), but due to the nature of the study, a number of analytic rating scales were modified and used for rating the essays.

Rating Scale: The rating scale used in this study contains seven dimensions: (1) Content, (2) Organization, (3) Vocabulary, (4) Mechanics, (5) Language Use and Grammar, (6) Formal Register, and (7) Fluency. These dimensions were adapted from Bachman and Palmer (1996), Kondo-Brown (2002), Lee (2002), Matsuno (2009), and Schaefer (2008). Considering the fact that the rating scale should allow the raters to exercise their judgment on as many factors as possible to constitute the construct of writing ability, the researchers combined the five mentioned rating scales in order to have a more comprehensive scale concerning the underlying constructs of the writing ability (see Appendix I). Accordingly, content was adapted and defined based on Kondo-Brown (2002), Matsuno (2009), and Schaefer (2008); Organization based on Lee (2002), Bachman and Palmer (1996), and Matsuno (2009); Vocabulary and Mechanics according to Kondo-Brown (2002); Language use and grammar based on Kondo-Brown (2002) and Schaefer (2008); Register according to Bachman and Palmer (1996); and Fluency based on Schaefer (2008).

The seven-dimension scale was first piloted on a sample of 36 sophomore English students similar to that of the main study. According to the results of the pilot study and regarding the opinions of some experienced instructors of writing concerning the weightings of the dimensions in above-mentioned sources, different weightings determined the scoring of each dimension in the rating scale. Thus, content was rated on a scale of 0-4 point, organization, vocabulary, mechanics and use on a scale of 0-3 point and fluency and register on a scale of 0-2 point. In order to make the dimensions of the rating scale distinct for the raters, they were provided with detailed description of them.

Paying attention to the fact that the dimensions of a rating scale should all be related to and measure a single construct (the writing ability) and also should not have any overlap in measuring the constituents of that construct, the Facets analysis of the rating scale dimensions was conducted. Masters' (1982) partial credit model (PCM) was used for data analysis. Of course, Rasch and IRT models can accommodate both dichotomous and polytomous scoring. According to Masters (1982), when items are scored dichotomously (i.e., right and wrong) the dichotomous model of Rasch (1960) can be used to model responses. However, when items are scored on a rating scale with more than two categories (Likert items), the dichotomous Rasch model cannot be used. In such cases, we need polytomous models such as Andrich's (1978) rating scale model (RSM) or Masters' (1982) partial credit model (PCM). These are extensions of the dichotomous Rasch model for polytomous items. RSM assumes that number of categories in all items is the same. But PCM doesn't have this restriction and can accommodate items with different number of response categories. Since different writing dimensions have different weightings in the rating scale, PCM was used for the analysis.

As Table 2 shows, there is no overfitting dimension which means that all the seven dimensions are necessary and each adds unique information to capture the overall writing ability of the test-takers. In addition, there is no underfitting dimension which means that all the seven dimensions fit the Rasch model and form a unidimensional writing scale. Thus, all the dimensions work together and the ratings on the individual dimensions can be added to come up with a single summary score to report examinees' writing performance. This is encouraging and suggests that this rating scale can be used as a reliable tool in the assessment of the learners' writing ability.

Rating Scale Proper Functioning Analysis: In order to investigate the proper functioning of the rating scale, rating scale indexes for each dimension were studied. Table 1 shows the rating scale statistics for the seven dimensions. The first column Dim shows the dimension score; column 2 shows the number of times that dimension or the score is observed in the data; the value in parentheses shows the percentage of the count. Column 3 shows the mean of the examinees who are scored on the dimension. We expect average measures to increase with category values. Column 4 shows the model expected value for column 3 (ie., the model predicted measure of the examinee's ability if the data fitted the Rasch model

perfectly). If the observed and expected examinee ability measures are close, the outfit mean square is close to its ideal value of 1. The larger the discrepancy between observed and expected measures, the larger the mean square index will be. An outfit mean square greater than 2 for a dimension suggests that ratings in that category for some examinees may not be contributing to meaningful measurement of the variable (Linacre, 1999).

Table 1

Summary of rating scale diagnostics for all dimensions

	Dim	Count(%)	Average Measure	Expected Measure	Outfit MnSq	Threshold	Threshold Error
Vocabulary	1	44(24)	-1.61	-1.56	.9	-	-
	2	102(55)	-.18	-.16	1	-1.75	.20
	3	40(22)	1.96	-1.86	.9	1.75	.23
Mechanics	0	2(1)	-.22	-.89	1.3	-	-
	1	61(33)	.26	.15	1.2	-3.8	.73
	2	86(43)	1.44	1.55	1.1	.47	.19
	3	37(20)	3.51	3.47	.9	3.34	.24
Content	1	3(2)	-.26	-.53	1.1	-	-
	2	40(22)	.47	.36	1.9	-2.68	.61
	3	73(39)	1.38	1.51	1.9	.30	.21
	4	70(38)	3.37	3.31	.8	2.38	.20
Organization	0	2(1)	-.03	-.42	1.2	-	-
	1	19(10)	.66	.44	1.2	-2.26	.74
	2	96(52)	1.45	1.64	.7	-.62	.26
	3	69(37)	3.78	3.58	.8	2.87	.19
Use	0	13(7)	-1.59	-1.94	1.2	-	-
	1	94(51)	-.85	-.76	1.0	-3.36	.31
	2	60(32)	.83	.85	.8	.45	.19
	3	19(10)	2.88	2.63	.9	2.91	.30
Register	0	2(1)	.21	.20	1.3	-	-
	1	121(65)	1.59	1.59	.8	-3.25	.72
	2	63(34)	3.71	3.71	1.1	3.25	.19
Fluency	0	10(5)	.23	-.39	1.2	-	-
	1	102(55)	.79	.87	.8	-2.12	.35
	2	74(40)	2.90	2.87	1.0	2.12	.19

Thresholds show whether the dimensions on the rating scales differentiate between high and low proficiency examinees. We expect the thresholds to be

reasonably widely separated along the proficiency continuum. Threshold estimates show the distinctiveness of each step on a Likert scale. Thresholds which are very close show that raters cannot distinguish among the dimensions. Thresholds which are too far from each other indicate that the number of dimensions on the scale is not enough and we need more to avoid loss of information. We expect thresholds to increase with category values.

As Table 1 shows, the rating scales used for each dimension function properly and average measures and thresholds advance with dimension scores and all fit the model.

Procedure

Thirty-two essays written by Iranian male and female university students aged 20 to 27 were collected at the Islamic Azad University, Quchan Branch. To control the possible topic-type effect, the students were all given the same topic (A happening in my childhood). This topic was chosen because it required no special knowledge on the part of the students and seemed relevant and accessible to the sample. Six raters, three Ph.D. students majoring in English Teaching and three MA graduates of the same field, all trained to use the researcher's rating scale, rated these essays. The raters were all faculty members of Islamic Azad University, Quchan Branch and all had similar experience, including years of teaching experience and teaching writing courses in Iranian EFL context. The training session was approximately 30 minutes. First, the researcher explained the purpose of the study and the instruments to the raters. Then, they were instructed to follow the rating scale while rating the essays. All the raters rated the essays and the data concerning the learners' total score as well as their scores on the rating scale dimensions were put into analysis.

Data Analysis

The gathered data provided a wide range of possible analyses for addressing the aims of the investigation. All the 32 compositions were rated by the six raters on all the seven dimensions of the rating scale.

Three facets of examinee ability, rater harshness, and dimension difficulty were specified. The analyses were carried out by Facets version 3.67.1 Linacre (2010). The contribution of Many-Faceted Rasch Model (Linacre, 1989) is that it adds other facets to the previous two-faceted Rasch models; that is, raters can be added

to the measurement model to study and cancel out their effect on measurement. In judged performance, we can argue that apart from learner's ability and item difficulty, raters' leniency or harshness, and interactions among raters and learners, criteria, etc… also play an important role in measurement. So, Facets model which can include other facets of measurement and is not limited to persons' abilities and items' difficulties was developed to address this issue. Masters' model is not appropriate for this kind of analysis.

Results

As the results of the analysis, based on the difficulty of the seven dimensions, indicate (Table 2), the easiest dimension was register with a difficulty estimate of -.95 and the hardest one was vocabulary with a difficulty measure of 1.40."The acceptable range for infit mean square is 0.50 as a lower-control limit and 1.50 as an upper-control limit for the outfit mean-square statistic" (Linacre, 2010, as cited in Eckes, 2011, p. 421). "Other researchers suggested using a narrower range defined by a lower-control limit of 0.70 (or 0.75) and an upper-control limit of 1.30" (Bond & Fox, 2007; Smith, 2004, as cited in Eckes, 2011, p. 421). According to Linacre (2010), values smaller than 0.50 indicate overfit and values greater than 1.50 indicate underfit. As the table shows, there are no overfitting dimensions (infit mean square <0.50) and no underfitting dimensions (infit mean square >1.50). Of course, the dimension content shows misfit (outfit> 1.50). Since the purpose of this study is not scale validation and refinement and it is merely interested in examining the interactions between raters' elements and other facets of the rating design, the researchers did not adopt any strategy to deal with the misfitting elements. Misfitting element can "reveal valuable insights into assessor behavior (Esfandiari & Myford, 2013). Moreover, Linacre (2011) states that mean square values in the range of 1.5-2 are "unproductive for construction of measurement, but not degrading" (p.248). That is, mean square values above 1.5 and below 2 are not threats to measurement.

Table 2
Dimensions Measurement Report

Dimension	Difficulty	Error	InfitMnSq (Measure)	OutfitMnSq (Model S.E.)
Vocabulary	1.40	.15	.93	.93
Use	1.32	.13	.94	.95
Mechanics	-.10	.14	1.10	1.08
Fluency	- .25	.16	1.05	.98
Content	-.56	.13	.98	1.67
Organization	-.87	.14	.94	.87
Register	-.95	.19	1.01	1.03

Figure 1 presents the examinee ability, rater harshness and dimension difficulty measures on the interval logit scale.

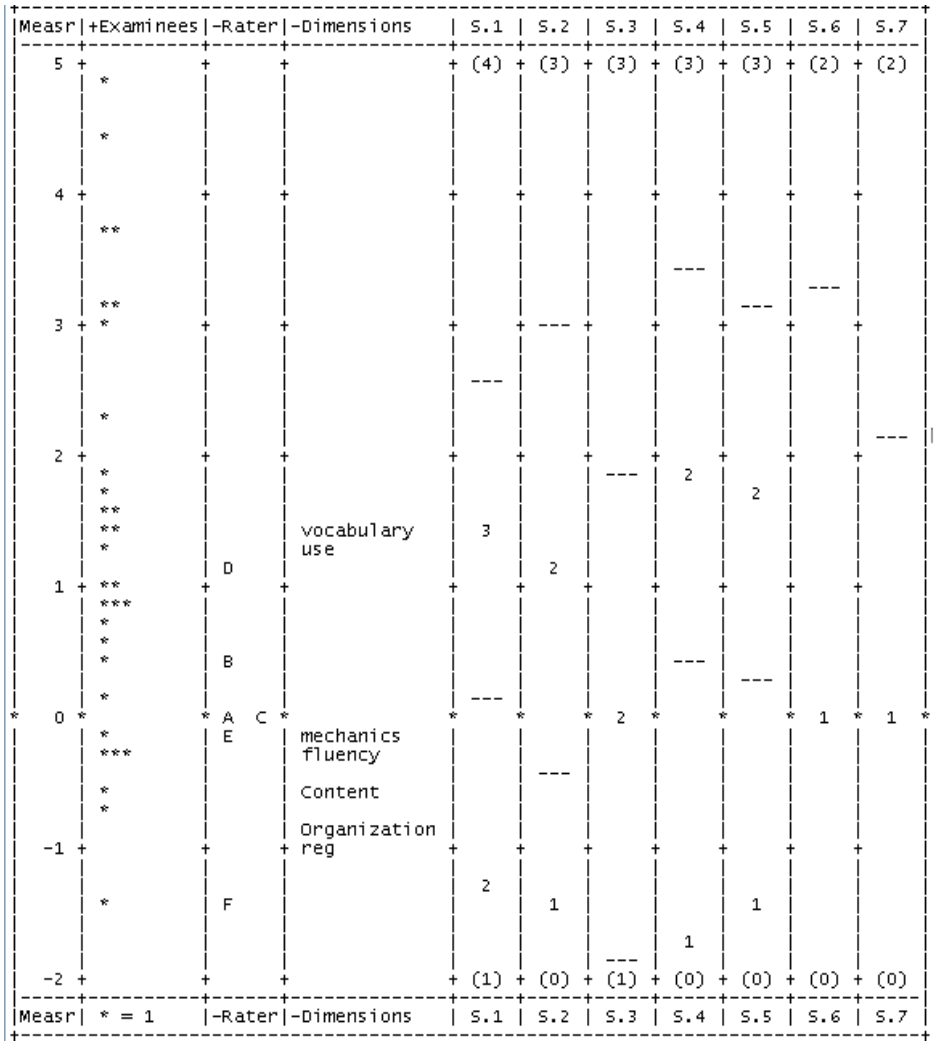

Figure 1: The examinee ability, rater harshness and dimension difficulty measures

According to Figure 1, vocabulary is the most harshly scored and register is the most leniently scored dimension. Columns 5 through 11 represent the rating scales

used to rate the examinees on each of the seven dimensions. The horizontal lines across each column show the point on the logit scale where the likelihood of getting the next higher score exceeds the likelihood of getting the next lower scoring for a given dimension (Myford, Marr, & Linacre, 1996). In other words, this predicts the likelihood of getting the specified score on a given scale sub-category for the examinees. For example, examinees with ability measures of below -2 to 0.30 are more likely to receive a rating of 2 on the first dimension (i.e. content) and those with an ability of 0.30 to 2.38 are more likely to receive a rating of 3 on this dimension.

Table 3 shows a lot of variation in examinees' ability ranging from -1.47 to 4.92 logits. The separation reliability of examinees' ability measures was .96 which indicates that the analysis reliably separates the examinees into different levels of ability (Separation=4.87, Strata=6.82). The chi square of 664.3 with 30 df is significant at $p<.001$; therefore, it is concluded that there is a lot of variation in examinees' writing ability. Regarding the acceptable range for infit mean square (lower-control limit of 0.50 and an upper-control limit of 1.50), only two examinees (24 & 12) with infit mean square values of 1.71 and 0.48 misfit which shows that the pattern of ratings of these examinees were inconsistent. This might be due to the fact that since examinee 24 has the highest ability measure, the raters' expectations and attitudes toward him might to some extent explain their inconsistent ratings of his examinee's writing task. Regarding examinee 12, the infit mean square value is 0.48 which is approximately equal to the lower control limit (0.50) and might be caused due to some unpredictable factors.

Table 3
Examinees Measurement Report

Examinee	Ability	Error (Measure)	InfitMnSq	OutfitMnSq (Model S.E.)
24	4.92	.46	1.71	4.27
11	4.37	.40	1.01	.68
25	3.69	.34	.84	.73
16	3.69	.34	.88	.68
20	3.15	.32	.70	.69
6	3.15	.32	1.05	.97
21	3.05	.31	1.12	1.00
8	2.24	.29	1.09	1.15
17	1.90	.29	.90	.90
1	1.64	.29	.91	.87
30	1.56	.29	.84	.84
26	1.56	.29	1.21	1.15
9	1.48	.29	1.00	.99
29	1.39	.29	.64	.64
7	1.31	.29	1.23	1.27
27	1.06	.29	1.02	1.02
5	.97	.29	.66	.69
28	.89	.29	1.18	1.18
15	.89	.29	.94	.92
3	.89	.29	1.02	1.03
12	.71	.29	.48	.42
19	.54	.30	1.09	1.14
32	.37	.30	1.27	1.28
10	.10	.30	1.11	1.10
13	-.08	.30	.64	.59
23	-.26	.30	1.38	1.21
22	-.26	.30	1.16	1.17
18	-.35	.30	1.36	1.57
4	-.63	.30	.82	.74
2	-.72	.31	1.19	1.24
31	-1.47	.31	.64	.67
Separation Reliability .96				

Regarding the first research question, 'To what degree do the raters differ from each other in their assessments of EFL learners' writing ability?', Table 4 shows that rater severity spanned between -1.49 (the most lenient) and 1.11 (the harshest). This is 2.60 logits of difference between the harshest and the most lenient rater which is an unacceptable high divergence. The reliability of the rater separation which shows the degree to which the analysis distinguishes between different levels of rater harshness is .97 (Separation=5.70, Strata=7.94). The chi square of 192 with 5 df is significant at $p<.001$, and this is an indicator that the raters consistently differ from each other in terms of overall harshness. In other words, there were significant differences among raters in their level of harshness. Meanwhile, all the raters' fit indexes are within the acceptable range of 0.50 -1.50 (except rater 2); that is, all raters were self-consistent in their ratings, except for Rater 2 (outfit>1.50) who has shown a little inconsistency in her rating.

Table 4
Rater Measurement Report

Rater	Sex	Education	Harshness (Measure)	Error	InfitMnSq (Model S.E.)	OutfitMnSq
4	Male	Ph.D. student	1.11	.13	.86	.85
2	Female	MA	.41	.13	1.3	1.63
1	Female	Ph.D. student	.07	.13	.93	.96
3	Female	Ph.D. student	.02	.13	.86	.85
5	Male	MA	-.11	.13	.76	.69
6	Female	MA	-1.49	.14	1.14	1.26
Reliability of the Rater Separation .97						

In order to answer the second research question, 'Does the interaction between raters and examinees cause bias in raters' assessment of EFL learners' writing ability?', the bias analysis was conducted.

Table 5 shows the results of bias analysis and the interaction of examinees and raters. In this table, the third column shows the score a certain examinee has been given by a certain rater. Observed Count is the number of dimensions. Obs-Exp Average shows the average observed-expected difference score from the given rater across the seven dimensions.

Table 5
Rater-examinee Interaction

Rater	Examinee	Observed Score	Expected Score	Observed count	Obs-Exp Average	Bias size	Model S. E.	T
2	7	7	12.5	7	- .78	- 3.04	.75	- 4.05
1	27	8	12.6	7	- .66	- 2.56	.75	- 3.39
6	23	9	13.1	7	- .59	- 2.23	.76	-.2.95
2	2	5	8.8	7	- .54	- 2.19	.80	- 2.74
1	10	7	10.8	7	- .55	-2.17	.75	- 2.88
2	22	6	9.6	7	- .51	-2.04	.76	- 2.68
5	15	9	12.7	7	- .51	-1.99	.76	- 2.63
1	26	10	13.7	7	- .53	-1.92	.75	- 2.56
4	21	11	14.6	7	- .52	-1.82	.74	-2.47
2	1	10	13.2	7	- .45	-1.67	.75	- 2.22
4	11	14	17.2	7	- .45	- 1.62	.69	- 2.34
6	18	10	13.0	7	- .42	- 1.57	.75	- 2.09
4	1	9	11.8	7	- .40	- 1.54	.76	- 2.03
2	26	16	13.0	7	.43	1.46	.72	2.04
2	10	13	10.2	7	.40	1.47	.70	2.09
1	22	13	10.2	7	.40	1.49	.70	2.13
1	23	13	10.2	7	.40	1.49	.70	2.13
3	1	17	14.0	7	.43	1.54	.76	2.01
4	18	11	8.2	7	.40	1.59	.74	2.15
1	32	15	11.3	7	.53	1.83	.70	2.63
2	27	16	12.0	7	.57	1.97	.72	2.75
1	18	14	10.0	7	.57	2.07	.69	2.99
2	28	16	11.7	7	.62	2.14	.72	2.99
2	8	19	14.4	7	.66	2.92	1.11	2.62

Bias size is the translation of Obs-Exp Average into logit units. Model S.E. is the error of the bias estimate. The letter *t* shows the statistical significance of the bias size, t values greater than 2 and lower than -2 are considered significant (A two-tailed 95% confidence interval is ±2 S.E. wide). As an example, the first row of the table shows the interaction between Rater 2 and Examinee 7. Rater 2 has given examinee 7 a score of 7. However, the model expects a score of 12.5 based on the overall ratings. This translates to 3.04 logit bias with a t of - 4.5 which is statistically significant. That is, this rater has scored this examinee harsher than

expected and as Table 5 additionally shows there are 24 cases of bias due to the rater-examinee interaction.

With regard to the third research question, 'Does the interaction between raters and rating scale cause bias in raters' assessment of EFL learners' writing ability?', Table 6 shows the results of bias analysis concerning the raters- rating scale dimension interaction.

Table 6

Raters- Rating Scale Category Interaction

Rater	Dimension	Observed Score	Expected Score	Observed count	Obs-Exp Average	Bias size	Model S. E.	T	Prob.
6	Fluency	43	50.9	31	-.25	- 1.30	.40	-3.28	.002
5	Register	36	41.4	31	-.18	-1.23	.50	-2.46	.019
3	Register	36	40.8	31	-.16	-1.10	.50	-2.21	.034
4	Register	32	36.0	31	-.13	-1.04	.51	-2.05	.049
5	Mechanics	49	58.3	31	-.30	-1.03	.34	-3.05	.004
2	Use	33	41.3	31	-.27	-.95	.34	-2.77	.009
1	Content	87	96.6	31	-.31	-.88	.30	-2.94	.006
1	Use	51	44.4	31	.21	.70	.32	2.18	.037
3	Content	105	97.2	31	.25	.78	.33	2.39	.023
4	Mechanics	55	47.4	31	.24	.85	.33	2.57	.015
1	Register	45	40.6	31	.14	.87	.43	2.01	.053
1	Fluency	47	41.3	31	.18	.90	.40	2.24	.032
2	Register	50	39.0	31	.35	2.11	.43	4.90	.000
chi-square: 133.6		df.: 42	significance (probability): .00						

As it can be seen in the first row, rater 6 has scored Dimension 7 (fluency) 43. This is the sum of all the scores which this rater has given to fluency over all 32 examinees. However, the model expects a score of 50.9. In other words, Rater 6 has scored fluency harsher than the model expects. This difference between observed score and model expected score translates to a bias of 1.30 logits, which is statistically significant since the *t* is lower than -2 and the probability of this difference occurring by chance alone is .002. Therefore, there is an interaction between Rater 6 and fluency and as Table 6 shows there are 13 cases of bias due to rater-rating scale interaction.

With regard to the fourth question 'Are there any *systematic* bias patterns due to rater-rating scale or rater- examinee facets among raters?', there exists no overall systematic bias pattern among the six raters concerning the bias resulting from

rater-rating scale interaction. However, some systematic bias pattern can be observed due to the rater-examinee interaction, since most of the bias cases deal with learners of extreme high or low ability. Thus, as it can be observed in Table 5, among 24 cases of bias due to rater-examinee interaction, the first 13 cases show the raters have scored the examinees harsher than expected and the last 11 cases show the raters' leniency toward the examinees. Of course, most of these bias cases deal with the examinees of extreme high or low ability.

Regarding the fifth research question 'Does the raters' rating differ from each other regarding rating scale category characteristics?', Table 7 shows the interaction between raters and dimensions in a pairwise fashion. Column 1 shows dimensions. Column 2, with three sub-columns, shows the statistics related to raters in relation to dimensions recorded in Column 1. Column 3, with three sub-columns, shows the statistics related to raters in relation to dimensions recorded in Column 1. In fact, in Columns 2 and 3, the ratings of pairs of raters are compared in relation to dimensions. In the first row of Table 7, Rater 2 and Rater 5 are compared in relation to register as one of the dimensions. Rater 2 perceives register to have a difficulty of -3.06 with a precision of .43 logits while rater 5 perceives this dimension to have a difficulty of .28 logits with a precision of .50 logits. The difference in their perception of the difficulty of register is 3.34 logits, which is statistically significant. In other words, raters 2 and 5 have different perceptions of the difficulty of register and don't have a common view of its difficulty.

Table 7
Rater-Dimension Interaction (Pairwise comparison)

Target Dimension	Rater	Target Measure	S.E.	Rater	Target Measure	S.E.	Contrast	t	Prob.
Register	2	-3.06	.43	5	.28	.50	-3.34	-5.07	.000
Register	2	-3.06	.43	3	.15	.50	-3.21	-4.88	.000
Register	2	-3.06	.43	4	.09	.51	-3.15	-4.74	.000
Register	2	-3.06	.43	6	-.80	.43	-2.26	-3.74	.000
Fluency	1	-1.15	.40	6	1.04	.40	-2.20	-3.90	.000
Register	1	-1.81	.43	5	.28	.50	-2.09	-3.18	.002
Register	1	-1.81	.43	3	.15	.50	-1.97	-2.99	.004
Register	1	-1.81	.43	4	.09	.51	-1.90	-2.86	.005
Mechanics	4	-.95	.33	5	.93	.34	-1.89	-3.98	.000
Use	1	.62	.32	2	2.27	.34	-1.65	-3.51	.000
Fluency	2	-.54	.39	6	1.04	.40	-1.59	-2.76	.006
Fluency	5	-.50	.40	6	1.04	.40	-1.54	-2.75	.007
Fluency	4	-.48	.39	6	1.04	.40	-1.52	-3.02	.007
Content	3	1.34	.33	4	.00	.30	-1.34	-2.33	.003
Fluency	1	-1.15	.40	3	.15	.39	-1.31	-2.78	.023
Use	1	.62	.32	6	1.88	.32	-1.26	-2.26	.007

Discussion

The first research question deals with the degree that raters differ from each other in their assessments of EFL learners' writing ability. As the data analysis shows, while all the raters were self-consistent across the dimensions in their ratings, they consistently differ from each other in terms of overall harshness and each rater has a different bias pattern. This finding is in accordance with Wigglesworth (1994), Lumley (2002), Kondo-Brown (2002), Eckes (2005), and Haiyang (2010) who have concluded that raters differed strongly in their severity with which they rated examinees. This shows that although all the raters were trained to use a single rating scale with modified dimensions, the subjective nature of writing assessment has caused significant differences among them in their level of harshness. This is an indicator of rater bias which is also confirmed by the results of bias analysis (Tables 5 & 6).

Regarding the existence of bias due to rater-examinee interaction, the bias analysis provides information about how well the performance of each individual rater matches the expected values predicted by the model generated in the analysis

(Sudweeks, Reeve, & Bradshaw, 2005). As the results show (Tables 5 & 6), rater bias has resulted from two kinds of interaction. The first is the interaction between raters and examinees. As Table 5 indicates, the interaction between some given raters and examinees has caused 24 cases of bias(out of 192) among which 13 cases show that the given examinee has been scored harsher than expected by the model while other 11 ones indicate that the given examinee has been scored more leniently. Meanwhile, most of these bias cases deal with the examinees of extreme high or low ability (18 cases). Furthermore, raters 1 (7 cases) and 2 (9 cases) are the most inconsistent raters, while the most consistent ones are raters 3 (1 case) and 5 (1 case). This finding is in accordance with what Kondo-Brown (2002) concluded in assessing U.S. university students' Japanese L2 compositions. The fact that rater–candidate bias interaction was much higher for candidates of extreme high or low ability might be to some extent the result of raters' expectations or attitudes concerning these groups of learners. That is, raters might expect high performance of candidates of higher ability, and therefore, they might rate their essays more severely while they might ignore the errors of the candidates of lower ability in order to encourage their performance. However, the results of this study show that rater-examinee interactions are observed specially for raters 1 and 2 and this indicates the need for rater training and awareness with regard to bias-related factors.

The next research question examines the interaction between raters and rating scale dimensions. As Table 6 shows, there are 13 cases of bias (out of 36) among which 5 cases deal with register, 2 fluency, 2 use, 2 content, and 2 mechanics. No rater-dimension interaction is observed concerning vocabulary and organization which shows that all the raters were consistent in assessing these two dimensions. The least consistent rater is rater 1 while the most consistent one is rater 6. Regarding all cases of bias (rater-examinee interaction and rater-dimension interaction), we see that nearly 50 percent of the cases deal with raters 1 and 2 (22 cases). Such evidence could suggest the fact that most of the raters' inconsistencies result from their own personal characteristics. Since in this study, the number of the raters is not so large that we can claim their individual differences (sex, education, etc) as the justifying cause of their harshness or inconsistent ratings, it would nevertheless be valuable to replicate this study with a large number of raters to investigate factors causing the raters' inconsistent ratings. Of course, it is evident that to some extent these inconsistencies can be removed through rater training and feedback provision.

Regarding the question 'Are there any systematic bias patterns due to rater-rating scale or rater-examinee facets among raters?', some systematic bias patterns were found among the six raters. This finding is in agreement with the results achieved by Kondo-Brown (2002) in assessing Japanese L2 compositions and Schaefer (2008) in investigating rater bias patterns in an EFL context. Since the percentage of significant rater–examinee bias interaction is much higher for learners of extreme high or low ability (learners of extreme high or low ability have been scored harsher or more leniently than expected), it can be suggested that raters are more likely to show more severe or lenient bias patters towards the highest or the lowest ability learners. Of course, the underlying factors of this tendency cannot precisely be predicted from this study, but this might be due to the raters' expectations and attitudes concerning these learners.

With regard to the question, 'Does the raters' rating differ from each other regarding the rating scale dimension characteristics?', as it was mentioned before, raters 2 and 5 have different perceptions of the difficulty of register and don't have a common view of its difficulty. This is also true for raters 2 and 3, raters 2 and 4, raters 1 and 5, raters 1 and 3, and raters 1 and 6 concerning register; while raters 1 and 6 have different views of the difficulty of fluency. In the same manner, as the findings (Table 7) indicate, all raters have different perceptions of the difficulty of the rating scale dimensions and don't have a common view of their difficulty which is in agreement with the results Kondo-Brown (2002) reported. Of course, this might be related to the raters' personal viewpoints concerning the importance of different components of the writing ability.

Conclusion

The results of this study confirm that the presence of factors other than learner's ability might influence the rates' judgments. Since many crucial decisions are taken based on these judgments, we are required to search for ways and strategies that guarantee fairer and more equitable judgments. The bias analysis is a valuable guide for this purpose since it will provide us with a certain amount of information about the probable factors underlying these unfair judgments.

The findings of bias analyses can be presented to the raters in order to make them aware of their biased tendencies toward learners or tasks, and as Wigglesworth (1993) indicated in a study of bias in oral proficiency, this feedback can improve the consistency of the raters' performance in subsequent ratings.

These results also reconfirm the need for rater training in performance assessment which, of course, has not yet been seriously regarded in Iranian educational settings. It is essential that rater trainers and test providers be aware of the raters' biased tendencies and their underlying causes so that they can provide more informative training sessions for the raters as well as more continuing feedback on their individual performances. This study was an attempt to show the need for such an issue and was limited in scope. Further studies can be conducted on different participants and in other environments in order to confirm or reject these findings. Such research will provide researchers with effective and informative findings to be applied in removing the problem of raters' inconsistent judgments and helping the students be accepted or rejected just based on their ability or knowledge of the trait intended by the examiners.

Notes on Contributors:

Mahnaz Saeidi *is Associate Professor of Applied Linguistics at Tabriz Branch, Islamic Azad University, Tabriz, Iran. She is the editorial board member of The Journal of Applied Linguistics. In 2007, 2008, 2009, 2010, and 2011 she won an award for being the best researcher. Her major research interests are multiple intelligences, focus on form, and assessment.*

Mandana Yousefi *is a Ph.D. graduate of TEFL at Islamic Azad University, Tabriz Branch, Tabriz, Iran. She is Assisstant Professor of Applied Linguistics at the English Department of the Islamic Azad University, Quchan, Iran. She has taught M.A. and B.A. courses at Quchan Islamic Azad University. Her areas of interest include Individual differences, writing instruction, and assessment.*

Purya Baghaei *is assistant professor in the English Department of Islamic Azad University, Mashhad Branch, Mashhad, Iran. His major research interests are the application of Rasch models in educational and psychological testing. His publications include extensive work on foreign language proficiency testing and scale development.*

References

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. New York: Oxford University Press.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah NJ: LEA.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications. Retrieved from http://books.google.com/books?id=kQOFIKjCILEC&pg=PA22&lpg=PA22&dq.

Congdon, P. (2006*). Bayesian statistical modeling* (2nd ed). England: John Wiley & Sons Ltd.

Deregowski, J. B., & Serpell, R. (1971). Performance on a sorting task: A cross-cultural Experiment. *International Journal of Psychology*, *6*, 273-281.

Eckes, T. (2005). Examining rater effects in Test DaF writing and speaking performance assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Eckes, T. (2011). Item banking for C-tests: a polytomous Rasch measuring approach. *Psychological Test and Assessment Modeling, 53*(4), 414-439.

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9,* 270-292.

Einarsdóttir, S., & Rounds, J. (2009).Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior: 74*, 295-307.

Engelhard, G. Jr. (1992). The measurement of writing ability with a Many-Faceted Rasch Model. *Applied Measurement in Education, 5*(3*)*, 171-191.

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessments for all students: Validity, technical adequacy, and implementation* (pp. 261-288). Mahwah, N. J.: Lawrence Erlbaum Associates.

Haiyang, S. (2010). An application of classical test theory and Many Facet Rasch Measurement in analyzing the Reliability of an English Test for Non-English Major Graduates. *Chinese Journal of Applied Linguistics*, *33* (2), 87-102.

Kondo-Brown, K. (2002). A FACET analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lee, Y. J. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, *8* (2), 135-157.

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

Linacre, J. M. (2010). *Facets Rasch Measurement computer program, version 3.67.1*. Chicago: Winsteps.com.

Linacre, J. M. (2011). *A user's guide to FACETS: Rasch-model computer programs* [Software manual], Chicago: Winsteps.com.

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., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1991). The invariance of judge severity calibrations. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago, IL.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch Measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

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

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.

McNamara, T. F., & Adams, R. J. (1991). Exploring rater behavior with Rasch techniques. *Paper presented at the 13th annual Language Testing Research Colloquium*, Princeton, NJ.

McNamara, T., & Roever, C. (2006). *Language testing: the social dimension*. Oxford, UK: Blackwell Publishing.

Moon, T. R., & Hughes, K. R. (2005). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21*(2), 15-19.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (TOEFL Research Rep. No. 95-40). Princeton, NJ: Educational testing Service.

Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgment processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation, 35,* 29-36.

O'Neill, T. R., & Lunz, M. E. (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*, New York.

Retrieved from www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED398284.

Pawlikowska–Smith, G. (2002). *Canadian Language Benchmarks 2000: Theoretical Framework.* Centre for Canadian Benchmarks. Retrieved from www.language.ca.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research*. Chicago: The University of Chicago Press.

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

Shultz, S. K., & Whitney, D. J. (2005). *Measurement theory in action: case studies and exercises.* NewDelhi: Sage Publications.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and Many-Facet Rasch Measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Van de vijver, F. J. R., & Poortinga,Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13,* 29-37.

Van deVijver, F., & Tanzar, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue européenne de psychologie appliquée, 54*, 119-135, Retrieved from www.sciencedirect.com

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

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 behavior in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, *17*(2), 77-103.

Appendix

The Modified Essay Rating Scale (MERS)

Category	Category
1. Content (4 points): a. complete knowledge of subject b. main idea is clear & well-developed c. interesting topic presentation d. thorough & sophisticated topic development e. topic relevance f. sufficiency of detail	**2. Organization (3 points):** a. overall organization introduction and thesis statement body and topic sentence development conclusion rhetorical control b. intersentential organization: ideas clearly supported textual cohesion and logical sequencing intersentential relationships appropriate use of transitions consistent style
3. Vocabulary (3 points): a. adequate range b. effective word/idiom choice c. word form master d. no meaning confusion e. no translation	**4. Mechanics (3 points):** a. spelling b. punctuation/ capitalization c. paragraph indentation d. handwriting e. paragraphing

5. Language use and grammar (3 points): a. correct and natural grammar b. well-constructed sentences c. balance of simple and complex sentences d. few errors of agreement, tense, number, word order, pronouns, inflections, …	**6. Formal register (2 points):** a. appropriate use of discourse markers and of formal register b. Sensitivity to register includes discourse in a specific subject matter (specialist or technical domain (e.g., the language of law); awareness of the differences between spoken and written mode of discourse; and use of style (e.g., frozen, formal, consultative, casual, and intimate.) (Joos, 1967, as cited in Pawlikowska–Smith, 2002). **7. Fluency (2 points):** a. length of the essay fulfills topic requirement b. sentences are sufficiently long

Adapted from Bachman and Palmer, 1996; Kondo-Brown, 2002; Lee, 2002; Matsuno, 2009; and Schaefer, 2008.