The Journal of Applied Linguistics and Applied Literature: Dynamics and Advances, Volume 9, Issue 1, Winter and Spring, 2021, pp. 93-119

Rater-Mediated Assessment of Iranian Undergraduate Students' College Essays: Many-Facet Rasch Modelling

Rajab Esfandiari
Associate Professor of Applied Linguistics, Department of English Language, Faculty of Humanities, Imam Khomeini International University, Qazvin, Iran
Email: esfandiari@hum.ikiu.ac.ir

Abstract

In rater-mediated assessments, the ratings awarded to language learners' written, or spoken, performances do not necessarily reflect their language abilities because a number of other construct-irrelevant factors may affect the knowledge they demonstrate. Rater subjectivity and rating scales are among the variables possibly influencing the final results. The purpose of the present study was to examine the extent to which university students' ratings on their essays mirrored the effect of these two factors. To that end, 150 Iranian EFL teachers rated ten five-paragraph essays BA students had written as their course requirements at Imam Khomeini International University. The raters used two rating scales to rate the essays on a number of assessment criteria. The study rested on a partial rating design, and the Rasch-based computer program, FACETS, was used to analyze the data. Results of Facets analyses showed raters differed considerably in the amounts of severity they exercised when rating the essays. The results also showed rater bias interactions with holistic rating scales. The implications of the findings for proposing procedures for reducing the effects of such extraneous variables are discussed.

Keywords: Analytic Scales, Bias, Holistic Scales, Rater Subjectivity, Severity

ARTICLE INFO
Research Article
Received: Saturday, December 12, 2020
Accepted: Friday, April 23, 2021
Published: Saturday, May 15, 2021
Available Online: Friday, April 23, 2021
DOI http://dx.doi.org/10.22049/jalda.2021.27032.1234

Online ISSN: 2383-2460; Print ISSN: 2383-591x

Introduction

In rater-mediated second language assessments, the ratings raters assign to test takers' performance result from the interaction of the rater characteristics with assessment criteria, rating scales, and testees' performance in testing settings. The rating process is a challenging task, during which rater subjectivity into the rating session may lead to construct-irrelevant variance (Wind, 2020). As Cronbach (1990) rightly asserted, when raters participate in the rating process, they are actually engaged in a "complex and error-prone cognitive process" (p. 584), in which they may not necessarily follow the rating criteria consistently and confidently, even after strict rater training (Knoch, Zhang, Elder, Flynn, Huisman, Woodward-Kron, Manias, & McNamara, 2020).

Rater variability is a blanket term researchers use to refer to the errors that raters introduce into the rating setting. Such errors, also known as rater effects (Myford & Wolfe, 2004), threaten the validity and fairness of decisions to be made about the future lives of test takers because they do not reflect the real abilities language learners demonstrate. Rater variability can manifest itself in a variety of ways. Raters may sometimes differ in the degrees of severity, or leniency, they exercise relative to assessment criteria and their colleagues in the rating process. They may also exhibit halo effect, central tendency effect, and restriction of range (Engelhard & Wind, 2017).

In addition to raters, rating scales exert an influence on ratings. As Myford and Wolfe (2004) neatly put it: "a rating scale is a measurement instrument used to record the results of the rater's observations" (p. 388). Similarly, Eckes (2015) noted that the type of rating scale makes a difference in helping language educators to make major assessment decisions. This decision is crucial because, as Weigle (2002) reminded us, "the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Rating scales may be problematic in the rating of essays. For instance, it may not be completely clear to raters what they are being asked to rate (Knoch, 2011). Similarly, McNamara (1996) noted that raters might have different interpretations of rating scales they were using. The rating scale categories might also be ambiguously worded, or two or more rating scale categories might overlap in meaning, blurring the boundaries between them (Weigle, 2002).

The interaction between raters and rating scales affects the findings. It is almost impossible to confidently claim the ratings awarded to language learners' written, or spoken, products accurately reflect their linguistic abilities because those ratings may also partly show raters' subjectivity. Language learners' performance may be a function of both their language abilities and raters' impressions. This runs the risk of making fair decisions about their placement into a higher level of language proficiency. Given the significance of accurate ratings, the present researcher set out to examine the subjectivity raters carry with them into rating sessions, estimating the extent to which their ratings were accurate indicators of language learners' written essays. Therefore, the following two research questions were formulated to achieve this major goal in this study.

1. To what extent are EFL teacher raters severe, or lenient, when rating EFL essays using rating scales?

2. In what ways are teacher raters different in severity, or leniency, in relation to assessment criteria?

Rater Variability in Second Language Performance Assessment

Recent studies have focused on identifying and conceptualizing different factors in second language performance assessment. In performance assessments, the testing setting is characterized by an interaction between the rater, the rating scale, rating processes, and the student performance (McNamara, 1996). A central issue in performance assessment is rater variability which is defined as the way(s) raters may introduce construct-irrelevant variance in the scores awarded to students' second language performance (Myford & Wolfe, 2003). Raters may show significant differences in the severity of their scoring as a result of their subjectivity, or inconsistency (Bonk & Ockey, 2003).

Eckes (2015) defined rater variability as the variability in the scores that can be attributed to raters rather than test takers. Since this variability has nothing to do with the test taker's language ability, it is considered a source of construct-irrelevant variance (Messick, 1989). Thus, rater variability may endanger the fairness and the validity of the scores assigned to test takers' performance. Messick identified two general threats to construct validity: construct underrepresentation and construct-irrelevant variance. In the former case, observations do not include all important aspects of the construct being measured. However, in the latter case, observations include aspects beyond the construct being measured. Both construct underrepresentation and construct-irrelevant variance change the interpretations of what the test claims to measure.

Rater variability may be related to many facets of the testing situation, such as the tasks (Wigglesworth, 1994), testing occasions (Lumley & McNamara, 1995), students / test takers (Kondo-Brown, 2002), and rubrics/scoring criteria (Eckes, 2005). Several empirical studies have focused on rater behavior in performance assessment and have revealed that rater behavior is closely related to unwanted rater variability, or inconsistency (Bachman, Lynch & Mason, 1995). Eckes (2008) aptly put that rater variability stems from the following sources: (a) the extent to which raters conform to the rating scale, (b) the way they use the rating scale criteria in rating sessions, (c) the extent to which raters exercise severity, or leniency, (d) the way they understand and apply rating scales, and (e) and the extent to which they rate consistently.

Many sources of variability can affect the scores awarded to test takers by raters other than writing ability in writing performance assessment. Such sources, as Schoonen (2005) pointed out, refer to the prompts test takers need to write about; various genres they are required to develop; imposed time restrictions; mode of delivery (paper-based, computer adaptive, or Internet-based); raters' inconsistent ratings; scoring procedures; and constructs to be rated.

As Myford and Wolfe (2003) noted, researchers studying rater performance address their work as investigations of "rater effects", "rater biases", and / or "rater errors" (p. 391), arguing that these terms are most often undifferentiated in the research literature and are usually used interchangeably. According to Scullen, Mount, and Goff (as cited in Myford and Wolfe, 2003, p. 391), rater effects are a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957). Many studies of second language performance assessment, including both second language writing and speaking, have demonstrated that rater errors tend to be systematic than random (McNamara, 1996; Upshur & Turner, 1999).

Myford and Wolfe (2003) stated that classic psychometric rater effects include severity / leniency, halo effect, central tendency, and restriction of range. Myford and Wolfe listed many other effects which, they claimed, are less frequently examined such as (1) inaccuracy, (2) logical error, (3) contrast error, (4) influences of rater biases, beliefs, attitudes, and personality characteristics, (5) influences of rater / ratee background characteristics, (6) proximity error, (7) recency (or primacy) error, (8) order effects, and (9) carryover effects (item-to-item carryover effects and test-to-test carryover effects). Because the primary focus of the present study is on severity and leniency, this concept is discussed in great detail in the following paragraphs

According to Cronbach (1990), severity / leniency effect is the most serious error that a rater can introduce into a rating setting. The term *leniency* was first coined by Kneeland (1929), who described this effect as the predisposition of a rater to "rate well above the midpoint of the scales used" (p. 356). Ford (as cited in Myford & Wolfe, 2003) coined the term *severity* to describe the other end of the continuum. Differences in severity between raters are exceedingly common (Eckes, 2015; McNamara, 1996) and also durable over time (Lim, 2012). The persistence of severity differences has prompted some researchers (Linacre, 2007) to argue for the employment of more sophisticated methodological breakthroughs such as many-facet Rash measurement to account for and model such differences.

According to Engelhard (1994), severity can be regarded as a continuum on which the raters range from lenient (one extreme of the continuum) to severe (the other extreme of the continuum). The reliability of separation index provides evidence of whether or not the raters systematically differ in severity. When different students / test takers are rated by different raters, the potential biasing effects of rater severity / leniency should be scrutinized.

Saal, Downey, and Lahey (1980) identified three different approaches for identifying this effect. One approach is to compare the mean of the ratings of the traits with the mid-points of the scales employed in the assessment. Another approach is to look for a significant rater effect within an analysis of variance (ANOVA) framework. The third approach is to explore the degree of skewedness of the frequency distributions of the ratings for the traits.

Myford and Wolfe (2003) provide the following strategies which have been offered in the literature in order to minimize the severity / leniency effect. The first strategy has to do with the definitions of the traits. Myford and Wolfe asserted that a clear specification of the definitions of the traits and providing crystal clear anchor descriptions for various scale categories offer the raters a perfect comprehension of what each scale means. Myford and Wolfe believe that, by doing so, there will be no ambiguities and the raters will be able to distinguish between the various levels of a specific trait. Devising rating scales that have several scale categories on the positive side and few scale categories on the negative side is the next strategy in order to counteract raters' tendencies to be lenient. Instructing and training raters to be aware of the severity / leniency effect is the third strategy. Asking raters to assign ratings using a forced distribution and having them place a pre-determined number of students / test takers in each rating category are other suggested techniques.

Rating Scales for Assessing Writing

Rating scales, also known as rubrics, are defined as "a guide listing specific criteria for grading or scoring academic papers, projects, or tests, and an instrument that describes a specific level of performance within a scale" (Crusan, 2015, p. 1). In educational settings, teachers are always concerned with many issues regarding the rating systems. Teachers want to know what to weigh in their assessment of students' performance, equity, and fairness in the assessment process, and comparability of evaluation. They want to realize whether their evaluation of a given student's performance matches another teacher's appraisal (Crusan, 2010). To minimize the effects of such factors on test takers' test scores, assessors have decided to focus their attention on rating scales, or rubrics (Dempsey, PytlikZillig, & Bruning, 2009).

Researchers who support the employment of rubrics argue that such instruments provide us with reliable and accurate evaluation (Crusan, 2010). Additionally, rubrics result in the clarification of criteria, showing test takers how their performance is evaluated (Ferris & Hedgcock, 2014). Rating scales function as the de facto test construct, although as a simplification of the construct (North, 2003). This conceptualization of the rating scale and its position in almost any discussion on the issue of performance assessment are inextricably bound up with each other. Rating scales are in fact "realizations of theoretical constructs, of beliefs about what writing is and what matters about writing" (Hamp-Lyons, 2011, p.3).

Four types of rating scales are reported in L2 literature: primary-trait, multiple-trait, holistic, and analytic. In the present study, the last two scales are explained because only these two were used in the study. The holistic rating scale requires a general impression of writing. Hamp-Lyons (1991) defined a holistic rating scale as a scoring method in which "each reader of a piece of writing reads the text rather quickly (typically one minute or less per handwritten page) and assigns a single score for its writing quality" (p. 243). She also stated that "this may be done wholly subjectively, or (and more commonly nowadays) by reference to a scoring guide or rubric, in which case it is often known as known as 'focused holistic scoring'" (pp. 243-244). White (1985) remarked that holistic scoring "reinforces the vision of reading and writing as intensely individual activities involving the full self" and that any other approach would be "reductive" (p. 33).

Holistic rating has some advantages. Holistic scoring is fast and less expensive (Weigle, 2002). A second advantage is that the reader's attention is focused on the strengths of writing (White, 1985). It is also a valid method of scoring (Mousavi, 2012).However, holistic scoring is not without limitations and flaws. As far as theoretical issues are concerned, Weigle (2002) believes that the reliability of holistic rating scale is lower than that of an analytic one (but still acceptable). Moreover, as far as the impact is concerned, the holistic rating scale might mask an uneven writing profile because it provides just a single score which may be misleading, particularly for placement.

Another type of rating scale is analytic rating scale. Weigle (2002) remarked that "in analytic scoring, scripts are rated on several aspects of writing or criteria rather than given a single score" (p. 114). The criteria used to assess the scripts or essays may include grammar, content, organization, mechanics of writing, and many other aspects. These rating scales have been largely in language teaching and testing situations. Weigle (2002) summarized a number of advantages of an analytic rating scale. She asserted that providing more useful diagnostic information about the writing abilities is the first advantage of analytic rating. She also stated that "analytic scoring is more useful in rater training" (p. 120). Analytic scales typically provide higher reliability and have higher construct validity for second language writers (Weir, 2005).

Despite all the merits, this method has major shortcomings. Analytic scoring procedure is extremely time-consuming and expensive to construct (Weigle, 2002) because the learners' scripts have to be corrected several times depending on the number of criteria involved (Mousavi, 2012). The second shortcoming of analytic scoring is that a tremendous amount of useful information will be lost when scores on different categories have to be combined to make a composite score (Knoch, 2011).

Empirical studies investigating the reliability of analytic and holistic rating scales have revealed mixed results. Some studies show that holistic rating scales are typically most reliable when they are used by experienced raters (Barkaoui, 2010). Novice raters, on the other hand, do not often have a clear conception of language proficiency (Isaacs & Thomson, 2013) and may be much more strongly impacted by the communicative and argumentative quality of the performance than by its form. Therefore, analytic rating scales can provide the required explicit guidance for novice raters (Harsch & Martin, 2013).

Methodology

Participants

Two groups participated in this study: students and teacher raters. Forty-five male and female BA students studying English Translation and English Language Teaching in two intact Essay Writing classes at Imam Khomeini International University in Qazvin participated in the study. These students had already passed Advanced Writing. They were approximately within the same age range.

One hundred and fifty English as a Foreign Language (EFL) Iranian non-native English speaking teachers in two language institutes participated in this study. The teacher raters were both female and male (one hundred and four female raters and forty-six male raters). They ranged in age from 24 to 60. They came from five language backgrounds: one hundred and thirty-seven teacher raters were native-Farsi speakers (91.33%), ten teacher raters were native-Turkish speakers (6.67%), and one teacher rater was native-Armenian speaker, one native-Maazani speaker, and one native-Kurdish speaker. Twenty-four of them (16%) had experience living in an English-speaking country. They had taught writing courses from 1 to 30 years. They were 39 BA (26%), 98 MA (65.3%), and 13 PhD (8.7%) holders in English Language Teaching, English Literature, Translation Studies, and other fields.

Data Collection Methods

Two assessment instruments were used in this study: students' essays and rating scales. Forty-five 5-paragraph essays collected from undergraduate (BA) students were used in the study. The students enrolled in two Essay Writing courses at Imam Khomeini International University in Qazvin, Iran. The students in Essay Writing classes were taught punctuation, the necessity of indentation, expression, features of a well-written 5-paragraph essay such as organization, content, transitions and coherence as well as principles for writing one-paragraph and 5-paragraph essays. The students in these classes were also taught various patterns of development, including comparison and contrast essays, cause-and-effect essays, and enumeration essays. After eight weekly meetings (each lasting one hour and a half), the instructor of the course told his students that they would take the midterm exam the following week.

During the exam, students had 90 minutes to write an expository comparison-contrast, 5-paragraph essay ranging in length from 500 to 700 words on the following topic: *An e-mail and a letter are both used to transfer information. There are, however, some differences between these two communication systems. Discuss three differences between them.* Ten essays were randomly selected for teacher raters to rate: essays 1, 5, 10, 15, 20, 25, 30, 35, 40, and 45. Then, the essays 1, 5, 15, 25, and 35 were assigned for analytic rating, and essays 10, 20, 30, 40, and 45 were assigned for holistic rating.

An analytic rating scale was developed based on the ESL Composition

Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981), the Composition Grading Scale (Bailey & Brown, 1984, as cited in Farhady, Jafarpour, & Birjandi (1994), and the principles of the comparison and contrast five-paragraph essay, to rate expository essays in the current study. The main criteria in the analytic rating scale were: (1) Organization, (2) Content, (3) Mechanics, (4) Grammar, (5) Vocabulary, and (6) Coherence and Transitions. Each criterion consisted of five descriptors. The scale categories ranged from 1 (very poor), 2 (poor), 3 (good), 4 (very good), 5 (excellent) (see appendix A).

A holistic rating scale was used to rate expository essays in this study. In the analytic scale, the six assessment criteria included five categories each. In designing a holistic rating scale, five categories were also used to score the essays. The scale categories ranged from 1 (very poor), 2 (poor), 3 (good), 4 (very good), 5 (excellent). Each category had its own distinctive descriptors (see appendix B).

Procedures

In order to fulfill the objectives of the present study, certain steps were followed. The researcher managed to conduct training sessions for raters before the rating process. The rater trainer was the researcher himself. He held a thirty-minute training session for each and every of the teacher raters and provided explanations on how to rate the essays analytically and holistically. Teacher raters rated expository essays, using analytic and holistic rating scales. The trainer also presented some previously rated expository essays rated both holistically and analytically to raters. In order to enhance the effectiveness of the training program, the trainer then asked raters to rate some essays individually holistically and analytically. In cases where raters assigned completely different ratings, they were asked to explain their highly unexpected ratings.

Expository essays written by the students were handed in to the teacher raters to rate both holistically and analytically based on what they acquired during the training session. The raters were asked to rate the first five essays based on the analytic rating scale and the second five essays based on the holistic rating scale. They were also asked to leave comments when necessary about various elements and features of the scripts and correct the students' errors if necessary. They were supposed to hand in the rated essays within two weeks. The raters received feedback on their assessment of the essays after the analyses of the data were completed.

Data Analysis

In this study, the following computer program was used to analyze the data. FACETS (version 3.68.1, Linacre, 2011) was used to ensure about the proper functioning of rating scales and to analyze the ratings for severity / leniency. It was also used to examine bias between teacher raters and the assessment criteria of the rating scales.

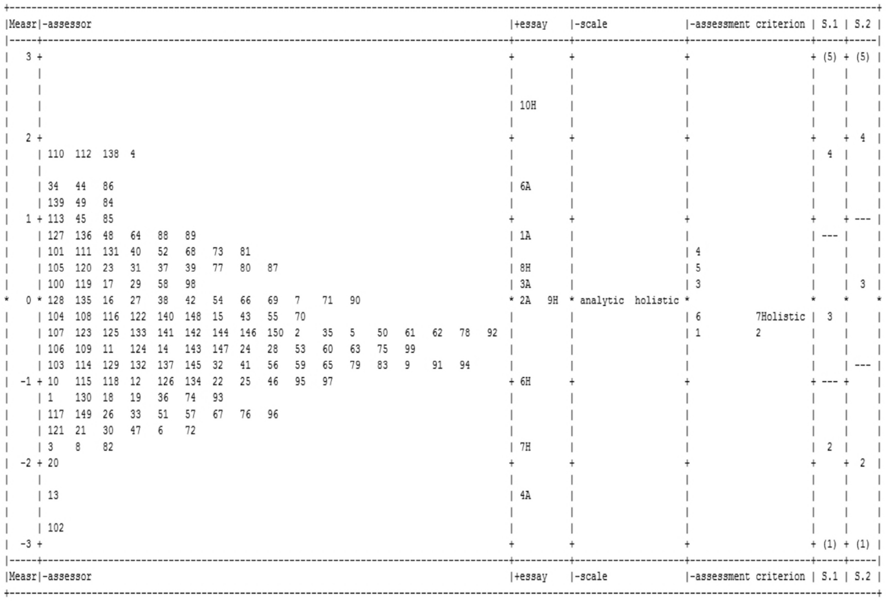
Results

The first research question of the present study was concerned with the extent to which EFL teacher raters exercised severity or leniency when rating EFL essays using rating scales. To answer this research question, the researcher used the following procedures.

The Vertical Rulers

Figure 1 reprsents vertical rulers (also known as Wright map and variable map) obtained from the Facets analysis in this study.The Rasch rating Andrich scale was used for the analysis of data in the curent study which included the following facets: rater, essay, scale, and assessment criterion. All these facets are portrayed in Figure 1.

Figure 1. Variable Map from FACETS



```
+-------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|-assessor                                                                           |+essay    |-scale              |-assessment criterion | S.1 | S.2 |
|-----+------------------------------------------------------------------------------------+----------+--------------------+----------------------+-----+-----|
|   3 +                                                                                    +          +                    +                      + (5) + (5) |
|     |                                                                                    |          |                    |                      |     |     |
|     |                                                                                    |          |                    |                      |     |     |
|     |                                                                                    | 10H      |                    |                      |     |     |
|     |                                                                                    |          |                    |                      |     |     |
|   2 +                                                                                    +          +                    +                      +     +  4  |
|     | 110  112  138  4                                                                   |          |                    |                      |  4  |     |
|     |                                                                                    |          |                    |                      |     |     |
|     | 34   44   86                                                                       | 6A       |                    |                      |     |     |
|     | 139  49   84                                                                       |          |                    |                      |     |     |
|   1 + 113  45   85                                                                       +          +                    +                      +     + --- |
|     | 127  136  48   64   88   89                                                        | 1A       |                    |                      | --- |     |
|     | 101  111  131  40   52   68   73   81                                              |          |                    | 4                    |     |     |
|     | 105  120  23   31   37   39   77   80   87                                         | 8H       |                    | 5                    |     |     |
|     | 100  119  17   29   58   98                                                        | 3A       |                    | 3                    |     |  3  |
*   0 * 128  135  16   27   38   42   54   66   69   7    71   90                          * 2A   9H  * analytic  holistic *                      *     *     *
|     | 104  108  116  122  140  148  15   43   55   70                                    |          |                    | 6          7Holistic |  3  |     |
|     | 107  123  125  133  141  142  144  146  150  2    35   5    50   61   62   78   92 |          |                    | 1          2         |     |     |
|     | 106  109  11   124  14   143  147  24   28   53   60   63   75   99                |          |                    |                      |     |     |
|     | 103  114  129  132  137  145  32   41   56   59   65   79   83   9    91   94      |          |                    |                      |     | --- |
|  -1 + 10   115  118  12   126  134  22   25   46   95   97                               + 6H       +                    +                      + --- +     |
|     | 1    130  18   19   36   74   93                                                   |          |                    |                      |     |     |
|     | 117  149  26   33   51   57   67   76   96                                         |          |                    |                      |     |     |
|     | 121  21   30   47   6    72                                                        |          |                    |                      |     |     |
|     | 3    8    82                                                                       | 7H       |                    |                      |  2  |     |
|  -2 + 20                                                                                 +          +                    +                      +     +  2  |
|     |                                                                                    |          |                    |                      |     |     |
|     | 13                                                                                 | 4A       |                    |                      |     |     |
|     |                                                                                    |          |                    |                      |     |     |
|     | 102                                                                                |          |                    |                      |     |     |
|  -3 +                                                                                    +          +                    +                      + (1) + (1) |
|-----+------------------------------------------------------------------------------------+----------+--------------------+----------------------+-----+-----|
|Measr|-assessor                                                                           |+essay    |-scale              |-assessment criterion | S.1 | S.2 |
+-------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Thus, it provides a very informative and unique frame of reference. The first column indisplays *the logit scale*, which ranges from -3 to +3 logits. The second column portrays *the assessors* (i.e., the severity levels exercised by the raters who rated the students' essays). Raters higher on the map rated the essays severely; by contrast, raters lower on the map rated the essays leniently.The third column displays *the essays*. The essays above the mean received low ratings while those below the mean received high ratings. The fourth column displays *the rating scales*. The fifth column displays *the assessment criteria.* Assessment criteria 3, 4, and 5 were difficult for students to receive high ratings on; by contrast, assessment criteria 1, 2, 6, and 7 were easy for students to receive high ratings on.

Proper Functioning of the Rating Scales

One important issue in the context of rater-mediated writing performance assessment concerns the quality of the rating scales that assessors employ to rate the students' essays. Two rating scales were used in this study: (1) an analytic rating scale, and (2) a holistic rating scale. In the following two sections, effectiveness of these two scales is investigated.

The category statistics (Table 1) provide useful pieces of information about the effectiveness of the *analytic rating scale* used in this study. According to Linacre (2004), in order for a rating scale to function effectively, a number of requirements should be met: (a) there should be at least 10 observations, or ratings, in each rating scale category; (b) average measures should advance monotonically with scale categories; (c) outfit mean-square values should be less than 2; (d) there should be monotonic advance for step difficulties (or scale calibrations); and (e) there should be more than 1.4, but less than 5, logits, for step difficulties (or scale calibrations).

Linacre (2004) pointed out that the existence of at least 10 observations in each rating scale category ensures achieving accurate threshold calibrations. The second column demonstrates that the first requirement is met. The second step is controlling the quality of the average measures. Average measures should advance monotonically with categories. It means that the higher the category, the larger the average measure.According to Table (1), the recorded average measure are: -2.44, -1.15, 0.29, 1.23, and 2.02 for categories 1 to 5 respectively. The average measures advance monotonically with categories; thus, it is safe to conclude that the second requirement is met.

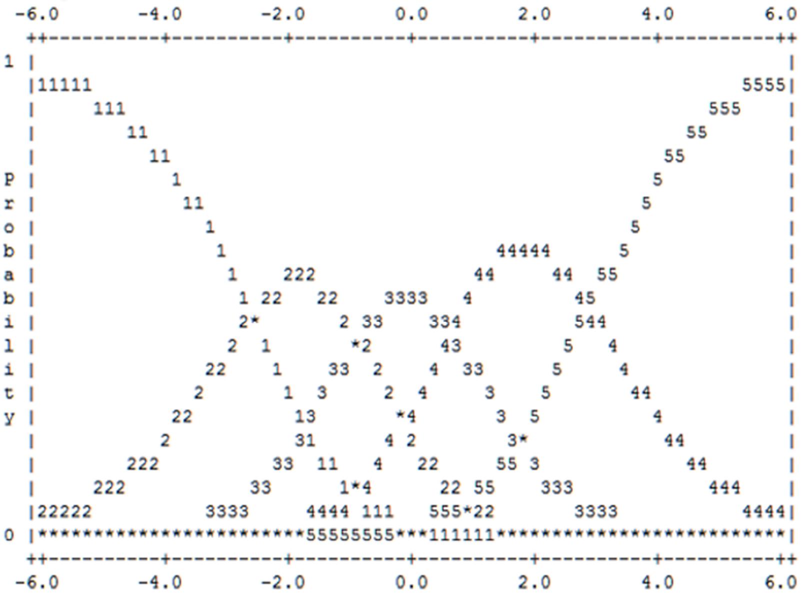
The third important indicator of rating scale effectiveness refers to the mean-square outfit statistic which is computed for each rating category. The mean-square outfit statistic makes a comparison between the average measures and the expected measures for each category. The ideal value for this indicator is 1, and this statistical indicator should not exceed 2.00. In the present study, the computed mean-square outfit statistic values are: 1.1, 0.9, 0.9, 0.9, and 1.0 for scale categories 1 to 5, respectively. Thus, the values are equal, or very close to the expected value of 1.

Another indicator of the quality of the rating scale is the ordering of the category thresholds. The thresholds should advance monotonically with categories ("the most probable from" column). According to Linacre (2004), step difficulties (or scale calibrations) should increase by 1.4, but less than 5 logits. According to Table 1, this requirement is not met in this study. Fortunately, as Linacre (2004) noted, "this degree of rating scale refinement is usually not required in order for valid and inferentially useful measures to be constructed from rating scale observations" (p. 274).

Table 1. Category Statistics for the Analytic Rating Scale

Categories	Counts	Average Measure	Outfit MnSq	Most Probable from
1	352	-2.44	1.1	low
2	802	-1.15	.9	-2.63
3	1323	.29	.9	-.88
4	1588	1.23	.9	.60
5	435	2.02	1.0	2.90

Figure 2 schematically illustrates the functionality of analytic rating scale. The figure, specifically, shows the category probability curves for the five-category rating scale the raters used when rating the students' essays on the six criteria (organization, content, mechanics, grammar, vocabulary, coherence and transitions). As can be observed in the probability curves, each category has a separate *peak*, which means that each category is the most probable.



```
 -6.0      -4.0       -2.0       0.0        2.0       4.0        6.0
  ++---------+----------+----------+---------+----------+---------++
1 |                                                                |
  |11111                                                       5555|
  |    111                                                  555    |
  |       11                                              55       |
  |         11                                          55         |
P |           1                                        5           |
r |            11                                     5            |
o |              1                                   5             |
b |               1                       44444     5              |
a |                1    222             44     44 55               |
b |                 1 22   22   3333   4         45                |
i |                 2*      2 33    334          544               |
l |                2  1      *2      43         5  4               |
i |              22    1    33 2    4  33      5    4              |
t |             2       1  3    2  4     3    5      44            |
y |           22         13      *4       3  5         4           |
  |          2           31     4 2        3*           44         |
  |       222          33  11  4   22     55 3            44       |
  |    222           33     1*4      22 55    333           444    |
  |22222         3333    4444 111   555*22       3333          4444|
0 |***********************5555555***111111*************************|
  ++---------+----------+----------+---------+----------+---------++
 -6.0      -4.0       -2.0       0.0        2.0       4.0        6.0
```

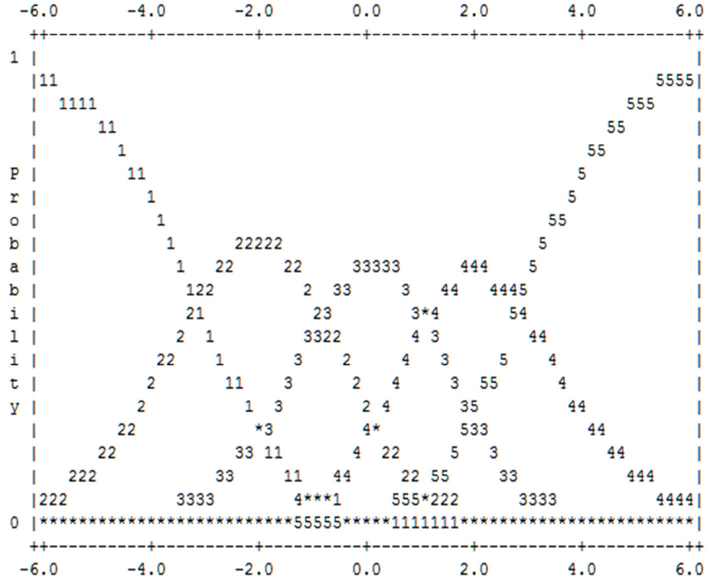
Figure 2. Category Probability Curves for Analytic Rating Scale

The same procedures, as explained in the previous subsection on the effectiveness of the analytic rating scale, were followed to ensure the proper functioning of the holistic rating scale. As Table 2 and Figure 3 show, this scale also functioned properly, meeting almost all the requirements outlined in the previous subsection

Table 2. Category Statistics for the Holistic Rating Scale

Categories	Counts	Average Measure	Outfit MnSq	Most probable from
1	30	-1.53	1.2	low
2	177	-.77	1.1	-3.12
3	266	.12	1.5	-.82
4	183	1.27	1.3	1.09
5	94	2.79	1.1	2.84

.



```
 -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
  ++----------+----------+----------+----------+----------+----------++
1 |                                                                   |
  |11                                                             5555|
  |  1111                                                      555    |
  |      11                                                  55       |
  |        1                                               55         |
P |         11                                            5           |
r |           1                                          5            |
o |            1                                       55             |
b |             1      22222                          5               |
a |              1   22     22     33333      444    5                |
b |               122         2  33     3   44   4445                 |
i |               21           23        3*4       54                 |
l |              2  1         3322       4 3         44               |
i |            22    1       3    2     4   3     5    4              |
t |           2       11    3      2   4     3  55      4             |
y |          2          1  3        2 4       35         44           |
  |        22            *3         4*        533          44         |
  |      22            33 11       4  22     5   3           44       |
  |   222            33     11   44     22 55     33           444    |
  |222           3333        4***1     555*222      3333          4444|
0 |**************************55555*****1111111************************|
  ++----------+----------+----------+----------+----------+----------++
 -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

Figure 3. Category Probability Curve for the Holistic Rating Scale

Table 3. displays the rater measurement report arranged by severity measures. The raters at the top of the table exercised more severity, while raters at the bottom of the table exercised more leniency in rating the students' essays. According to this Table, teacher raters can be divided into severe and lenient raters. Teacher raters with positive severity measures (measures above the mean) are severe, whereas teacher raters with negative severity measures (measures below the mean) are lenient. The rater measurement report reveals that 48 teacher raters in the present study were severe (with positive severity measures), while 102 teacher raters were lenient (with negative severity measures). The most severe rater had the severity measure of 1.89 logits and the most lenient rater had the severity measure of -2.77 logits. Thus, rater severity measures showed a 4.66-logit span.

Table 3. Rater Measurement Report

Rater	Measure	Model S. E.	Infit Mn Sq	Outfit Mn Sq
112	1.89	.25	1.14	1.15
4	1.77	.25	.72	.65
110	1.77	.25	1.63	1.63
138	1.71	.24	1.44	1.28
34	1.48	.24	1.52	1.44
44	1.42	.24	.96	.95
86	1.31	.24	.49	.51
49	1.14	.24	1.52	1.58
84	1.14	.24	2.04	1.86
139	1.14	.24	.93	.96
45	1.08	.24	1.40	1.47
113	1.08	.24	.76	.96
85	.97	.24	.60	.58
136	.86	.24	2.30	2.12
89	.80	.23	.84	.84
48	.75	.23	1.01	1.10
64	.75	.23	1.83	1.73
88	.75	.23	.89	.87
127	.75	.23	.39	.39
52	.69	.23	1.64	1.63
81	.69	.23	.50	.51
111	.69	.23	.81	.79
40	.64	.23	2.02	1.96
131	.58	.23	.77	.77
68	.53	.23	1.36	1.36
73	.53	.23	.84	.82
101	.53	.23	.78	.77
37	.47	.23	.93	.92
80	.42	.23	.71	.70

23	.36	.23	1.02	.99
39	.36	.23	.78	.80
105	.36	.23	2.79	2.71
120	.36	.23	.85	.87
31	.31	.23	.50	.51
77	.31	.23	.77	.80
87	.31	.23	.81	.80
17	.25	.23	.95	.96
29	.25	.23	.77	.80
119	.25	.23	1.35	1.35
58	.20	.23	.60	.64
98	.20	.23	.63	.64
100	.20	.23	.96	.93
90	.09	.24	1.20	1.33
7	.03	.24	1.29	1.25
27	.03	.24	.86	.87
38	.03	.24	1.13	1.14
66	.03	.24	.85	.89
128	.03	.24	1.08	1.07
16	-.02	.24	.67	.66
42	-.02	.24	.69	.69
54	-.02	.24	.70	.72
135	-.02	.24	.82	.82
69	-.08	.24	.76	.76
71	-.08	.24	.67	.66
15	-.14	.24	.80	.77
43	-.14	.24	.80	.77
108	-.19	.24	1.08	1.05
122	-.19	.24	.75	.77
140	-.19	.24	1.09	1.15
148	-.19	.24	.66	.66
55	-.25	.24	2.07	2.10
70	-.25	.24	1.05	1.05
104	-.25	.24	.82	.84
116	-.25	.24	1.47	1.39
35	-.30	.24	1.24	1.26
123	-.30	.24	1.48	1.45
150	-.30	.24	.68	.66
50	-.30	.24	.85	.87
62	-.36	.24	1.55	1.52
141	-.36	.24	.94	.94
142	-.36	.24	.95	.95
144	-.36	.24	.94	.94
2	-.36	.24	.99	1.01

5	-.42	.24	1.86	1.83
61	-.42	.24	.99	.98
78	-.42	.24	.30	.31
125	-.42	.24	.47	.46
133	-.42	.24	.95	.95
92	-.42	.24	.91	.89
107	-.48	.24	.69	.71
146	-.48	.24	.46	.45
14	-.48	.24	1.09	1.05
28	-.53	.24	.78	.79
53	-.53	.24	.67	.67
60	-.53	.24	1.31	1.34
63	-.53	.24	.62	.62
106	-.53	.24	.92	.91
124	-.53	.24	.47	.46
143	-.53	.24	1.20	1.15
147	-.53	.24	1.05	1.01
24	-.53	.24	.69	.68
11	-.59	.24	.97	.98
75	-.65	.24	1.12	1.11
99	-.65	.24	1.14	1.12
109	-.65	.24	.95	.90
56	-.65	.24	2.24	2.23
94	-.71	.24	.79	.80
103	-.71	.24	.95	1.01
114	-.71	.24	1.18	1.17
145	-.71	.24	.84	.84
9	-.71	.24	.75	.75
41	-.77	.24	1.64	1.63
59	-.77	.24	.46	.47
91	-.77	.24	.30	.30
132	-.77	.24	.52	.51
137	-.77	.24	.51	.53
83	-.77	.24	.87	.91
129	-.82	.24	.36	.34
32	-.82	.24	.86	.86
65	-.88	.24	1.36	1.39
79	-.88	.24	1.17	1.17
10	-.88	.24	1.02	1.00
12	-.94	.24	.43	.43
25	-.94	.24	.88	.88
95	-.94	.24	1.46	1.53
118	-.94	.24	.88	.89
126	-1.00	.25	1.15	1.11

22	-1.06	.25	.81	.82
46	-1.06	.25	1.01	1.04
97	-1.06	.25	1.35	1.34
115	-1.06	.25	.44	.45
134	-1.06	.25	.58	.60
18	-1.12	.25	1.29	1.24
19	-1.12	.25	1.70	1.57
74	-1.12	.25	2.01	1.97
93	-1.12	.25	1.64	1.77
1	-1.19	.25	1.13	1.15
130	-1.19	.25	.82	.83
36	-1.25	.25	.87	.87
33	-1.31	.25	.69	.69
76	-1.31	.25	.83	.82
149	-1.31	.25	.71	.67
26	-1.37	.25	.74	.77
57	-1.37	.25	.83	.85
67	-1.37	.25	.89	.88
51	-1.43	.25	1.07	1.04
117	-1.43	.25	.76	.77
96	-1.50	.25	1.16	1.12
121	-1.56	.25	.75	.79
21	-1.63	.25	1.11	1.10
6	-1.69	.26	.96	1.03
30	-1.69	.26	1.57	1.61
47	-1.69	.26	.75	1.08
72	-1.69	.26	.79	.82
3	-1.76	.26	.84	.84
8	-1.76	.26	1.16	1.21
82	-1.89	.26	.73	.72
20	-1.96	.26	1.26	1.68
13	-2.31	.27	.99	.97
102	-2.77	.29	.57	.57

The Second Research Question

The second research question of the present study addressed the extent to which EFL teacher raters differed in the degrees of bias towards assessment criteria. To answer this research question, the researcher used the following procedures.

The second research question concerns severity or leniency differences between the teacher assessors in relation to different assessment criteria.As mentioned previously, elements of different facets may interact with each other in the context of rater-mediated writing performance assessment. MFRM is particularly well suited to assessing these bias types and correcting for their impact (Bond & Fox, 2015). In the present study, the researcher investigated the possible interactions between assessors and assessment criteria as well as the interactions between assessors and rating scales. There were some statistically significant interactions between assessors and assessment criteria, and assessors and rating scales which are summarized in Table 4 and Table 5, respectively.

Table 4 portrays *the Assessor-Assessment Criterion Bias/ Interaction Analysis.* Values of *t*-score are important in making decisions about whether the interactions are statistically significant or not. The values of *t*-score smaller than "-2" or greater than "+2" are statistically significant. In Table 4, the holistic rating is considered a single criterion. Evidently, many assessors had significant interactions with the holistic criterion. Fifty interactions between assessors and assessment criteria were statistically significant, with *t*-scores either greater than +2 or smaller than -2. Twenty-five of the significant interactions are positive (showing severity), and twenty-five of the significant interactions are negative (showing leniency). Figure 4 presents a schematic representation of interactions between assessors and assessment criteria.

Table 4. Assessor-Assessment Criterion Bias / Interaction Analysis

Rater	Logit	Crit	Logit	Obs Score	Exp Score	Obs-Exp Ave.	Bias Size	Model S.E.	t score	Infit MnSq	Outfit MnSq
5	.42	1	-.38	11	17.3	-1.26	2.44	.63	3.85	2.0	2.2
44	-1.42	7	-.12	8	11.8	-.76	2.02	.83	2.44	.7	.6
20	1.96	7	-.12	15	19.6	-.93	1.99	.65	3.08	1.1	1.1
54	.02	7	-.12	11	15.1	-.83	1.80	.69	2.62	1.0	1.1
76	1.31	7	-.12	14	18.2	-.84	1.76	.65	2.71	.6	.6
109	.65	4	.54	11	15.6	-.92	1.75	.63	2.76	.7	.6
7	-.03	7	-.12	11	15.0	-.80	1.75	.69	2.54	2.0	1.8
88	-.75	4	.54	8	11.9	-.78	1.73	.74	2.33	.8	.7
30	1.69	7	-.12	15	19.1	-.81	1.73	.65	2.67	.2	.2
62	.36	7	-.12	12	15.9	-.79	1.68	.67	2.51	.4	.4
65	.88	3	.10	13	17.3	-.86	1.65	.61	2.70	.9	.9
92	.48	3	.10	12	16.3	-.85	1.63	.62	2.62	.2	.2

103	.71	7	-.12	13	16.8	-.75	1.59	.66	2.42	.5	.4
93	1.12	7	-.12	14	17.8	-.75	1.58	.65	2.43	1.0	1.0
133	.42	3	.10	12	16.1	-.83	1.57	.62	2.53	1.2	1.1
42	.02	3	.10	11	15.1	-.82	1.57	.63	2.47	.4	.3
29	-.25	7	-.12	11	14.5	-.70	1.53	.69	2.22	.3	.3
46	1.06	7	-.12	14	17.6	-.72	1.52	.65	2.34	.7	.7
15	.14	7	-.12	12	15.4	-.68	1.46	.67	2.18	.4	.4
43	.14	7	-.12	12	15.4	-.68	1.46	.67	2.18	.4	.4
53	.53	7	-.12	13	16.4	-.67	1.42	.66	2.15	.2	.2
116	.25	3	.10	12	15.7	-.74	1.40	.62	2.25	1.0	1.2
108	.19	2	-.34	13	16.7	-.73	1.40	.61	2.28	.2	.2
8	1.76	7	-.12	16	19.2	-.64	1.37	.65	2.12	.8	.8
149	1.31	7	-.12	15	18.2	-.64	1.34	.65	2.08	.2	.2
56	.71	7	-.12	20	16.8	.65	-1.42	.69	-2.07	1.9	1.6
108	.19	3	.10	19	15.5	.69	-1.43	.67	-2.13	.7	.7
58	-.20	7	-.12	18	14.6	.68	-1.43	.66	-2.17	.7	.6
99	.65	7	-.12	20	16.6	.67	-1.48	.69	-2.15	.1	.1
48	-.75	3	.10	17	13.1	.79	-1.51	.64	-2.37	.6	.6
103	.71	6	-.25	21	17.7	.66	-1.53	.73	-2.09	.9	.9
5	.42	5	.44	19	15.3	.75	-1.54	.67	-2.30	.3	.3
21	1.63	7	-.12	22	18.9	.62	-1.54	.77	-2.00	.6	.5
136	-.86	5	.44	16	11.9	.82	-1.56	.63	-2.50	2.2	2.2
107	.48	7	-.12	20	16.2	.76	-1.65	.69	-2.40	.1	.1
93	1.12	6	-.25	22	18.7	.66	-1.69	.80	-2.11	.9	1.2
141	.36	3	.10	20	16.0	.80	-1.72	.69	-2.48	.7	.7
144	.36	3	.10	20	16.0	.80	-1.72	.69	-2.48	.7	.7
10	.94	1	-.38	22	18.6	.69	-1.74	.80	-2.17	1.7	1.2

112	-1.89	7	-.12	15	10.8	.85	-1.86	.65	-2.87	1.6	1.5
109	.65	3	.10	21	16.7	.86	-1.94	.73	-2.65	.7	.9
52	-.69	6	-.25	19	14.1	.97	-1.96	.67	-2.93	.9	1.0
51	1.43	1	-.38	23	19.7	.67	-1.99	.94	-2.12	.7	.5
51	1.43	2	-.34	23	19.6	.69	-2.04	.94	-2.17	.7	.5
14	.53	7	-.12	21	16.4	.93	-2.09	.72	-2.91	1.0	2.5
68	-.53	7	-.12	19	13.8	1.03	-2.20	.67	-3.28	.1	.1
34	-1.48	7	-.12	17	11.7	1.07	-2.28	.65	-3.51	.6	.6
123	.30	7	-.12	21	15.8	1.04	-2.31	.72	-3.22	1.3	2.8
45	-1.08	7	-.12	20	12.5	1.49	-3.21	.69	-4.66	.5	.4
49	-1.14	7	-.12	20	12.4	1.52	-3.26	.69	-4.75	.5	.4

Fixed (all = 0) chi-square: 1080.6 d.f.: 1050 significance (probability): .25

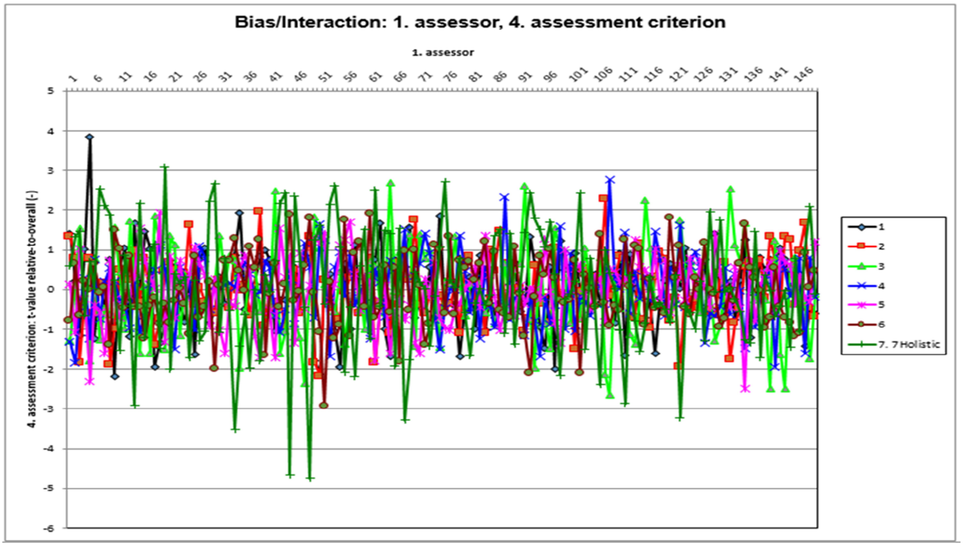


Figure 4. Bias Analysis for Assessment Criterion (Assessor-Items Interactions)

Note. Items: 1 = Organization, 2 = Content, 3 = Mechanics, 4 = Grammar, 5 = Vocabulary, 6 = Coherence and Transitions, 7 = Holistic

Table5 displays *the Assessor-Rating Scale Bias / Interaction Analysis*. The values of *t*-scores smaller than "-2" or greater than "+2" are statistically significant. Amazingly, the only statistically significant interactions between assessors and rating scales are between assessors and the holistic rating scale. All these assessors in Table 5 had significant interactions with the holistic scale. Twenty-eight interactions between assessors and the scale were statistically significant, with *t*-scores either greater than +2 or smaller than -2. Sixteen of the significant interactions are positive (showing severity), and eleven of the significant interactions are negative (showing leniency). Figure 5 presents a schematic representation of interactions between assessors and rating scales.

Table 5. Assessor-Rating Scale Bias / Interaction Analysis

Rater	Logit	Scale	Logit	Observed Score	Expected Score	Obs-Exp average	Bias Size	Model S E.	t score	Infit MnSq	Outfit MnSq
44	-1.42	2	.00	8	11.8	-.76	2.02	.83	2.44	.7	.6
20	1.96	2	.00	15	19.6	-.93	1.99	.65	3.08	1.1	1.1
54	.02	2	.00	11	15.1	-.83	1.80	.69	2.62	1.0	1.1
76	1.31	2	.00	14	18.2	-.84	1.76	.65	2.71	.6	.6
7	-.03	2	.00	11	15.0	-.80	1.75	.69	2.54	2.0	1.8
30	1.69	2	.00	15	19.1	-.81	1.73	.65	2.67	.2	.2
62	.36	2	.00	12	15.9	-.79	1.68	.67	2.51	.4	.4
103	.71	2	.00	13	16.8	-.75	1.59	.66	2.42	.5	.4
93	1.12	2	.00	14	17.8	-.75	1.58	.65	2.43	1.0	1.0
29	-.25	2	.00	11	14.5	-.70	1.53	.69	2.22	.3	.3

46	1.06	2	.00	14	17.6	-.72	1.52	.65	2.34	.7	.7
15	.14	2	.00	12	15.4	-.68	1.46	.67	2.18	.4	.4
43	.14	2	.00	12	15.4	-.68	1.46	.67	2.18	.4	.4
53	.53	2	.00	13	16.4	-.67	1.42	.66	2.15	.2	.2
8	1.76	2	.00	16	19.2	-.64	1.37	.65	2.12	.8	.8
149	1.31	2	.00	15	18.2	-.64	1.34	.65	2.08	.2	.2
56	.71	2	.00	20	16.8	.65	-1.42	.69	-2.07	1.9	1.6
58	-.20	2	.00	18	14.6	.68	-1.43	.66	-2.17	.7	.6
99	.65	2	.00	20	16.6	.67	-1.48	.69	-2.15	.1	.1
21	1.63	2	.00	22	18.9	.62	-1.54	.77	-2.00	.6	.5
107	.48	2	.00	20	16.2	.76	-1.65	.69	-2.40	.1	.1
112	-1.89	2	.00	15	10.8	.85	-1.86	.65	-2.87	1.6	1.5
14	.53	2	.00	21	16.4	.93	-2.09	.72	-2.91	1.0	2.5
68	-.53	2	.00	19	13.8	1.03	-2.20	.67	-3.28	.1	.1
34	-1.48	2	.00	17	11.7	1.07	-2.28	.65	-3.51	.6	.6
123	.30	2	.00	21	15.8	1.04	-2.31	.72	-3.22	1.3	2.8
45	-1.08	2	.00	20	12.5	1.49	-3.21	.69	-4.66	.5	.4
49	-1.14	2	.00	20	12.4	1.52	-3.26	.69	-4.75	.5	.4

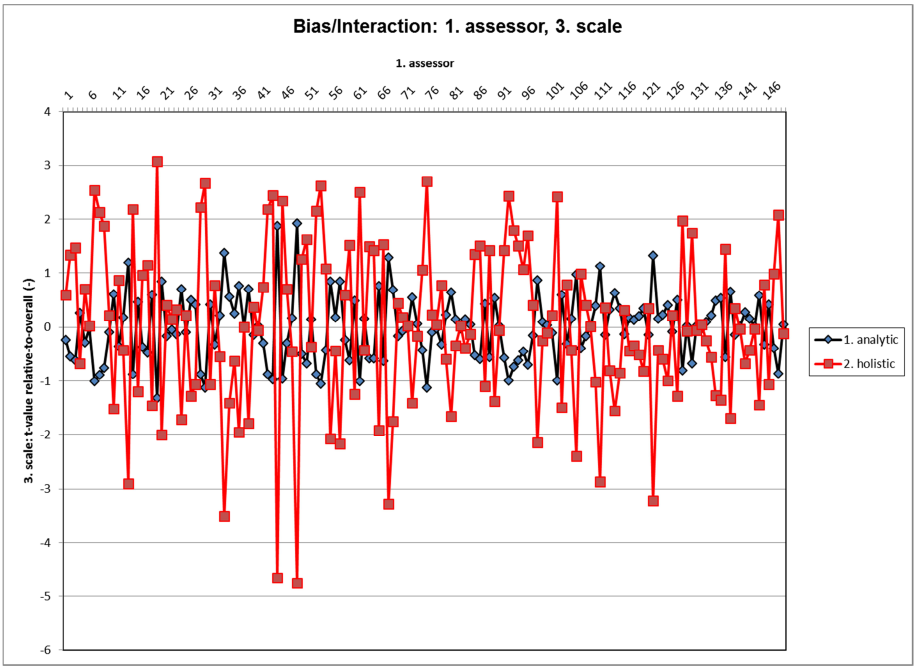


Figure 5. Bias / Interactions Between Assessors and Rating Scale

Note. 1: analytic rating scale, 2; holistic rating scale

Discussion

First of all, it was found that teacher raters did not function interchangeably, and they exercised varying degrees of severity. Based on MFRM output, the assessor severity measures had a spread of 4.66 logits between the most severe and the most lenient teacher assessors, and assessor separation index and assessor separation reliability were 3.49 and .92, respectively. Thus, assessor severity measures were far from homogenous. The finding that raters did not function interchangeably agreed well with several related lines of research on the levels of severity exercised by raters in language performance assessment (Bachman et al., 1995; Eckes, 2005; Kondo-Brown, 2002; McNamara, 1996; Myford & Wolfe, 2003). For example, Bachman et al. (1995) used GENOVA and FACETS to investigate variability in task and rater judgments in a performance test of foreign language speaking. They found that raters differed substantially in terms of the severity they had in their judgments. The most severe and least severe raters had measures of 1.93 and -2.27, respectively. Thus, the spread of severity measures was 4.20 for their study. Eckes (2005) investigated rater effects in the writing and speaking parts of TestDaf (Test of German as a Foreign Language). Using the MFRM approach in data analysis, Eckes found that raters differed strongly in the severity with which they assessed the test takers. Kondo-Brown (2002) used three raters, who were experienced Japanese language instructors at the same university to rate 234 essays. The severity span between the most lenient rater (Rater 1) and the most severe rater (Rater 2) was only .44 logits in Kondo-Brown's study. However, results of some studies have shown the partial efficacy of rater training in minimizing the errors that raters introduce into rating sessions (Elder et al., 2005; Knoch et al., 2007).

Another finding of the present study is that some raters had significant interactions with certain assessment criteria than others and these flagged interactions were not supported by the model expectations. There were fifty statistically significant interactions between teacher raters and assessment criteria. Twenty-five interactions were positive (having *t*-scores greater than +2, indicating severity) and twenty-five interactions were negative (having *t*-scores smaller than -2, indicating leniency). Out of the fifty statistically significant interactions, 28 interactions were between raters and the holistic (overall) criterion. Ten significant interactions were between raters and mechanics. Five significant interactions were between raters and content. There were also three significant interactions between raters and organization, two significant interactions between raters and grammar, and two significant interactions between raters and vocabulary.

The finding that different raters attach differential levels of importance to specific assessment criteria confirms the findings of some previous studies. For instance, in case of L2 writers, more importance is attached to discourse-level features (Kuiken & Vedder, 2014; Lee, 2009). Kuiken and Vedder (2014) reported that in case of L2 (two argumentative essays written by learners of Italian and of Dutch), raters had a tendency to give higher scores on communicative adequacy compared to linguistic complexity. Raters of both Italian and Dutch reported attaching more prominence to communicative adequacy (including, content, use of arguments, rhetorical organization, style and general comprehensibility) than to linguistic complexity (including, grammar, lexicon, spelling, and accuracy). The raters also stated that their expectations for lower level and higher level students differed, for both L2 and L1 writers. Raters did not point out any specific communicative, or linguistic, features when asked which features were associated with a particular rating level. Overall comprehensibility, clear text structure, and convincing arguments were crucial criteria for the lower proficiency levels, whereas the use of more complex syntactic structures and sophisticated words were considered to be more essential for the higher levels. Raters also remarked that they had more difficulty in assigning a text to either scale level 3 or 4, compared to the scale levels at the lower or upper end of the rating scale.

Since all teacher raters in the present study were Iranian non-native speakers of English and the most severely rated criterion was grammar, this finding can be endorsed by the study carried out by Marefat and Heydari (2016),who used both Iranian non-native teacher raters and native raters in their study and found that Iranian raters were more severe with grammar. This outcome of the present study has also been endorsed with by Hyland and Anan (2006), who gave a correction task to two groups of raters. The raters were Japanese and English EFL teachers. These two researchers found that NNES (nonnative English speaking) teachers exercised more severity when it came to grammatical errors. Furthermore, Lee (2009) reported that Korean raters find grammar as the most difficult criterion to score. However, Lee (2009) concluded that the greater stringency of NS (native speaking) teachers regarding the grammatical errors did not produce more accurate error correction.

Conclusions and Implications

The results of the present study, like many previous studies, indicated that the training provided for the teacher assessors could not eliminate differences in the severity levels that assessors exercised. Even though the results of some studies have indicated that rater training can reduce, but not necessarily eliminate, rater errors (Elder et al., 2005; Knoch et al., 2007), other studies exploring the effects of training on rater behavior have concluded that changing raters' severity levels, even with directive feedback, is a challenging enterprise (Knoch, 2011; Wigglesworth, 1993).

Since all the raters in this study were Iranian non-native English speakers, there is a need for rater training courses in the country. Many raters paid more attention to superficial and mechanical features of writing rather than to the communicative features. Furthermore, in case of holistic ratings, raters displayed more inconsistencies. However, as far as the rating scale is concerned, developing of a local rating scale that takes into account the particularities of the Iranian EFL assessment is highly appreciated. Such an objective measurement instrument mitigates any inconsistencies in the assessment.

Researchers and practitioners should find practical ways to ensure that the validity and fairness of the ratings will not be mitigated. The bias / interaction analysis serves as a practical source of information. The bias / interaction analysis provides researchers with information about possible variables underlying unfair assessments. According to Wigglesworth (1993), formative feedback can improve the consistency of the raters' performance in subsequent ratings. Thus, the findings of bias analyses can be presented to the teacher assessors to make them conscious of their biased predispositions toward assessment criteria or rating scales.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257. https://doi.org/10.1177/026553229501200206

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74. https://doi.org/10.1080/15434300903464418

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110. https://doi.org/10.1191/0265532203lt245oa

Cronbach, L. I. (1990). *Essentials of psychological testing* (5th ed.). Harper and Row.

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press.

Crusan, D. (2015). Dance, ten; looks: three: Why rubrics matter [Editorial]. *Assessing Writing, 26*(1),1–4. https://doi.org/10.1016/j.asw.2015.08.002

Dempsey, M. S., Pytlik Zillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38-61. https://doi.org/10.1016/j.asw.2008.12.003

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *5*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd edition). Frankfurt: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175-196. https://doi.org/10.1207/s15434311laq0203_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93-112.https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., & Wind, S. A. (2017). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge.

Farhady, H., Jafarpour, A., & Birjandi, P. (1994). *Testing language skills: From theory to practice*. The Organization for Researching and Composing University Textbooks in the Humanities (SAMT).

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). Routledge.

Hamp-Lyons, L. (1991). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-78). Cambridge University Press.

Hamp-Lyons, L. (2011). Writing assessment: Shifting Issues, new tools, enduring questions. *Assessing Writing*, *16*(1), 3–5. https://doi.org/10.1016/j.asw.2010.12.001

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 281-307. https://doi.org/10.1080/0969594X. 2012.742422

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System*, *34*(4), 509-519. https://doi.org/10.1016/j.system.2006.09.001

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

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

Kneeland, N. (1929). That lenient tendency in rating. *Personnel Journal, 7*, 356-366.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81-96.https://doi.org/10.1016/j.asw.2011.02.003

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Knoch, U., Zhang, B. Y., Elder, C., Flynn, F., Huisman, A., Woodward-Kron, R., Manias, E., & McNamara, T. (2020). I will go to my grave fighting for grammar: Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. *Assessing Writing, 46,* 1-14. https://doi.org/10.1016/j.asw.2020.100488

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329-348. https://doi.org/10.1177/0265532214526174

Lee, H. K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific Education Review*, *10*(3), 387-397.https://doi.org/10.1007/s12564-009-9030-3

Lim, G. S. (2012). Developing and validating a mark scheme for Writing. Cambridge ESOL: *Research Notes*, *49*, 6–9.

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

Linacre, J. M. (2007). Facets Rasch measurement computer program (Version 3.64*.2) [Computer software]*. Winsteps.com.

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

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing,* *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, *27*(1), 24-36. https://doi.org/10.1016/j.asw.2015.10.001

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed. (pp. 13–103). American Council on Education and Macmillan.

Mousavi, S. A. (2012). *An encyclopedic dictionary of language testing*. Rahnama Press.

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFLMonograph*, *24*(pp. 1-106).file:///C:/Users/RAJABE~1/AppData/Local/Temp/NORTHETS2003.pdf

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 82–11.https://doi.org/10.1177/026553229901600105

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave MacMillan.

White, E.M. (1985). *Teaching and assessing writing*. Jossey-Bass.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-335. https://doi.org/10.1177/026553229301000306

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77–103. https://doi.org/10.1075/aral.17.2.04wig

Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing, 43,* 1-14. https://doi.org/10.1016/j.asw.2019.100416

Authors' Biographies

Rajab Esfandiari is an Assistant Professor of Applied Linguistics at Imam Khomeini International University in Qazvin, Iran. His areas of interest and specialization include Teaching and Assessing L2 Writing, Multifaceted Rasch Measurement, L2 Classroom Assessment, and EAP Teaching and Testing. He can be reached via his email address:
Email: *esfandiari@hum.ikiu.ac.ir*