Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 16, No. 1; 2026, March

ISSN 2476-5880
https://www.ijlt.ir

A Comparison of the Multifaceted Rasch Model and Rating Scale Model for Evaluating EFL Writing Performance

Sareh Sahebalam[1], Purya Baghaei[2]*, Mojgan Rashtchi[3]

ARTICLE INFO

Article History:
Received: February 2025
Accepted: March 2025

KEYWORDS
Multifaceted Rasch model,
rating quality,
rating scale model,
writing assessment

ABSTRACT

This study compares the Multifaceted Rasch Model (MFRM) and the Rasch Rating Scale Model (RSM) in evaluating English as a Foreign Language (EFL) writing performance. Rater-mediated assessments depend on rating quality to ensure fairness and validity. The MFRM accounts for rater severity, task difficulty, and student ability, making it a more precise tool for performance-based evaluations. In contrast, RSM simplifies assessment by assuming all raters and tasks function similarly, thus failing to adjust for rater variability. The present study analyzed writing samples from 156 Iranian TEFL students, rated by five experienced IELTS instructors. Results showed that MFRM-adjusted student scores for rater severity, whereas RSM assigned identical scores to students with the same raw scores, ignoring rater effects. Item difficulty rankings were consistent across models, with language being the most difficult criterion and content the easiest. MFRM also revealed interactions between rater gender and student gender, though biases were statistically insignificant. While MFRM provides fairer, rater-invariant scoring, its computational complexity and limited accessibility remain challenging. RSM, though easier to implement, risks overlooking rater bias. The study concludes that MFRM is preferable for high-stakes and criterion-referenced assessments where fairness is crucial, whereas RSM may suffice for norm-referenced assessments. Future research should explore ways to simplify MFRM adoption, such as user-friendly software and training programs for educators.

1. Introduction

Assessments that rely on evaluators' judgments, known as rater-mediated assessments (Eckes, 2015; Engelhard, 2013), play a crucial role in various international evaluations, such as the Test of English as a Foreign Language (TOEFL) (Jamieson & Poonpon, 2013) and the International English Language Testing System (IELTS). A key factor in interpreting these assessments is the quality of ratings (Hamp-Lyons, 2007). Concerns regarding rating quality are widely discussed in research on performance-based assessments (Lane & Stone, 2006) and particularly in language assessment studies (McNamara, 1996). As a result, researchers have introduced multiple quality-control strategies for ratings, aligning with different measurement models. Since rating quality is fundamental to the psychometric validity of rater-mediated assessments, various evaluation methods reflect different perspectives on key elements, including both the quality of ratings and the raters themselves. For example, in a study by Taufiqulloh et al. (2025), it was emphasized that there is a significant need for more standardized and contextually valid assessment practices despite the complexity and diversity of

1 TEFL Department, Faculty of Foreign Languages, North Tehran Branch, Islamic Azad University, Tehran, Iran. Email: sahebalamsareh@gmail.com
2 English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran. Email: baghaei.purya@gmail.com
3 TEFL Department, Faculty of Foreign Languages, North Tehran Branch, Islamic Azad University, Tehran, Iran Email: rashtchi@gmail.com

writing assessments in EFL contexts. According to the study, effective assessments align with instructional goals and provide valuable feedback to enhance student learning. In another study, Shahi et al. (2025) suggested that conducting rater training prior to test administration could enhance the accuracy of subsequent assessments. Such training would emphasize consistent application of rating scales and establish a common understanding of the performance characteristics being evaluated by each test.

Polytomous Item Response Theory (IRT) models are considered valuable tools for examining different aspects of rater-mediated assessments. In language testing, IRT models—especially those based on Rasch models (Rasch, 1960)—are commonly used to analyze these assessments. Rasch models estimate the probability of a rating falling into a particular category based on the rater's level of severity and the student's demonstrated proficiency. These models also emphasize rater-invariant measurement, meaning that if achieved, student performance can be assessed independently of specific raters, and rater severity can be evaluated without being influenced by the students they score (Engelhard, 2013; Wright & Stone, 1979). Researchers applying Rasch-based measurement in rater-mediated assessments seek to determine the extent of rater-invariant measurement to refine evaluation methods and guide future research.

A systematic review by Wind and Peterson (2018) found that Rasch measurement theory is the most commonly used method for assessing ratings in language tests. The widespread adoption of Rasch-based techniques reflects longstanding support in language assessment research over the past forty years (McNamara & Knoch, 2012), particularly in second language (L2) testing (Eckes, 2005, 2015; McNamara, 1996). This popularity highlights the practical advantages of Rasch models in refining assessment methods. Unlike classical test theory, Rasch models provide rating quality metrics at the level of individual raters, measuring their adherence to invariant measurement criteria. Additionally, Rasch indices offer diagnostic insights into various rater-related factors, such as severity levels, rating scale usage, unexpected rating patterns, and other rater effects (Eckes, 2015; Engelhard, 2013).

1.1 Multifaceted Rasch Model

The Multifaceted Rasch Model (MFRM) is an extension of the Rasch measurement model, used to analyze data with multiple dimensions or "facets" that influence performance or responses. In traditional Rasch models, only one facet—typically the item difficulty—is considered in conjunction with a person's ability. The MFRM, however, allows for multiple facets to be included, making it a powerful tool for more complex scenarios where several variables impact outcomes. This model is frequently used in educational testing, psychological assessments, and performance evaluations.

In MFRM, each "facet" represents a different dimension that affects the responses. For example, in a language testing scenario, the facets might include: (1) person ability, (2) item difficulty, (3) rater severity, and (4) task characteristics (the complexity of the task or question). The model calculates the probability of a specific response based on these facets, estimating parameters for each while keeping them on the same measurement scale. In general, the MFRM formula is:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \theta_n - \delta_i - \rho_j - \tau_k$$

P_{nijk} and P_{nijk-1} are the probabilities of person *n* obtaining a score of *k* or *k−1* on item *i*, rated by rater *j*.
θ_n represents the ability level of the person *n*.
δ_i is the difficulty of the item.
ρ_j is the severity of the rater.
$\tau\kappa$ represents the threshold between adjacent score categories.

MFRM allows for the analysis of data with multiple influencing factors while maintaining a unidimensional scale. It can also adjust for rater severity or leniency, which helps to make the scoring fairer and reduces bias. MFRM estimates parameters for each facet independently, allowing for better insights into each component's role in the measurement. The MFRM is particularly useful in fields

where subjective judgments are common, such as performance assessments like art, music, or athletics where accounting for task difficulty and rater bias is paramount. The Multifaceted Rasch Model provides a robust approach for obtaining more accurate and fair measures when multiple facets influence scoring, making it ideal for complex assessments with multiple sources of variability.

The MFRM) is the most widely used Rasch-based approach for analyzing ratings in language testing (Eckes, 2015; Linacre, 1989; Wind & Peterson, 2018). Designed for performance assessments, MFRM independently models examinees, tasks, and raters, adjusting individuals' ability scores for variations in rater severity or leniency. However, the model is complex, and many researchers and practitioners are unfamiliar with its application. Additionally, user-friendly software for estimating MFRM parameters is limited.

Given these challenges, athuors, (under review) proposed using the more accessible Rasch Rating Scale Model (RSM) (Andrich, 1978) to analyze rater-mediated assessments. They argued that RSM effectively converts ordinal rating data into interval-level measurements, allowing for fair comparisons across individuals and items. It is particularly suitable for cases where all items share the same rating scale structure and aims to create a unified measurement scale that aligns person ability and item difficulty with the ordinal nature of response categories.

Athuors, (under review), also warned that RSM is not specifically designed for performance-based assessments and does not account for the rater facet, meaning it does not model raters or adjust ability scores for variations in rater severity. However, given the complexity of MFRM and the limited availability of user-friendly software, they aimed to explore the extent to which RSM can be effectively applied to evaluate rated performances.

In this study, we aim to compare RSM and the Mulufaceted Rasch model (MFRM) for evaluating writing assessments. In this endeavor, we rely on data from athuors, (under review).and analyze the same data set with MFRM, then compare the findings from the two analyses. The study will reveal if the RSM can be employed as a workable substitute for the MFRM.

2. Method

2.1 Participants

The study involved 167 Iranian TEFL students at the B.A. level from various universities in Mashhad. Of these, 110 were female, 28 were male, and 18 did not specify their gender. The participants were chosen through convenience sampling, and ultimately, 156 students submitted their essays. This sample size was sufficient to meet the requirements for RSM analysis (Linacre, 1994). Additionally, five raters participated in assessing the essays. All were non-native English speakers and experienced IELTS instructors with Master's or Doctoral degrees in TEFL or English literature. These evaluators had attained an overall IELTS band score of at least 8 and were either Ph.D. students in applied linguistics or TEFL. They also had a minimum of ten years of experience in teaching and evaluating writing skills.

2.2 Instruments

This study utilized two primary instruments: A writing prompt designed to elicit student responses and an analytical rating scale used to assess and score the essays. These tools were selected to systematically evaluate the participants' writing proficiency.

2.3 Writing Prompt

Participants were required to write an argumentative essay of at least 250 words within a 60-minute in-class session. The task aimed to assess their composition skills and ability to apply linguistic knowledge. The writing topic was selected from the Cambridge B2 First Level Writing Assessments and asked students to respond to the following prompt:

"Could you live without the internet for a month? Write and tell us what difference this would make to your life."

2.4 Rating Scale

Scoring rubrics are widely used in educational contexts for evaluating performance-based tasks. In this study, the Cambridge English Writing Assessment Scale (CEWAS, Cambridge English, 2024)

was employed to assess the essays. CEWAS is an analytic rating rubric that evaluates four key criteria—content, communicative achievement, organization, and language—using a five-point rating scale.

3. Analyses

3.1 Multifaceted Rasch Model

To compare MFRM and RSM, the rating data were first analyzed with the computer program FACETS (Version 3.71.4; Linacre, 2014). The program used the ratings that raters gave to students to estimate individual student abilities, rater severities, and scale category difficulties. Since the number of response categories was the same in all the rating criteria (content, communicative achievement, organization, and language) all four rating criteria had a common category structure. Hence, the specific model implemented in the analyses was a five-facet rating scale model (Linacre & Wright, 2002). All facets except the examinee facet were centered and the convergence criteria were left at their default values. The model was estimated and converged after 65 iterations. FACETS calibrates the examinees, raters, and criteria, (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis. Next, interaction effects, such as the interaction between raters and examinees or between raters and criteria, were detected by examining the standardized residuals (i.e., standardized differences between the observed and expected ratings). An interaction analysis (or bias analysis) helps to identify unusual interaction patterns among facet elements, particularly those patterns that point to consistent deviations from what is expected on the basis of the model.

3.1.1 Rater Severity and Fit

Figure 1 shows the variable map displaying the locations of the examinees, raters, and the rating criteria. Note that each star in the students' column represents two examinees, and a dot represents one. The horizontal dashed lines in column 7 indicate the category threshold measures. The figure also shows that raters' sex and students' sex were dummy variables and were fixed to zero.

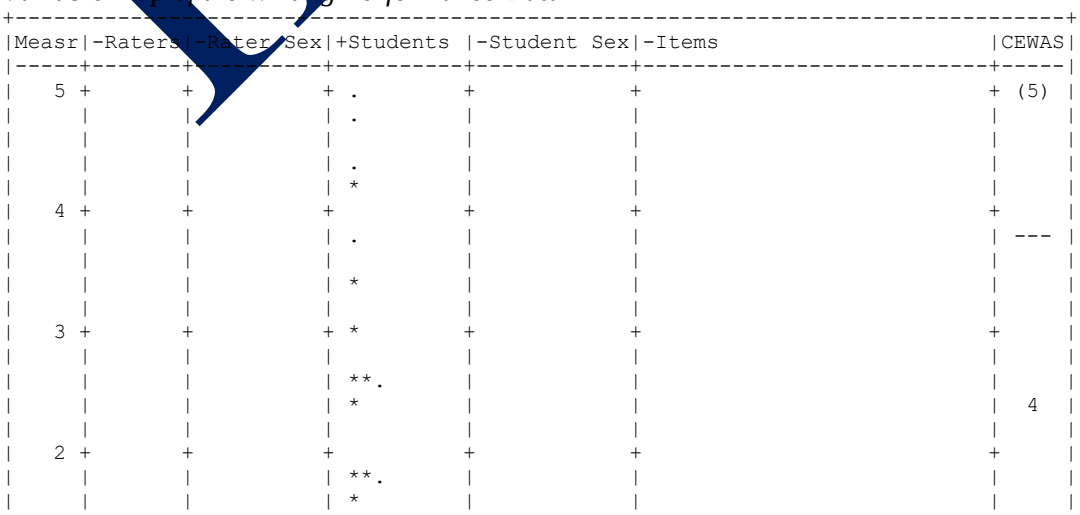
Table 1.

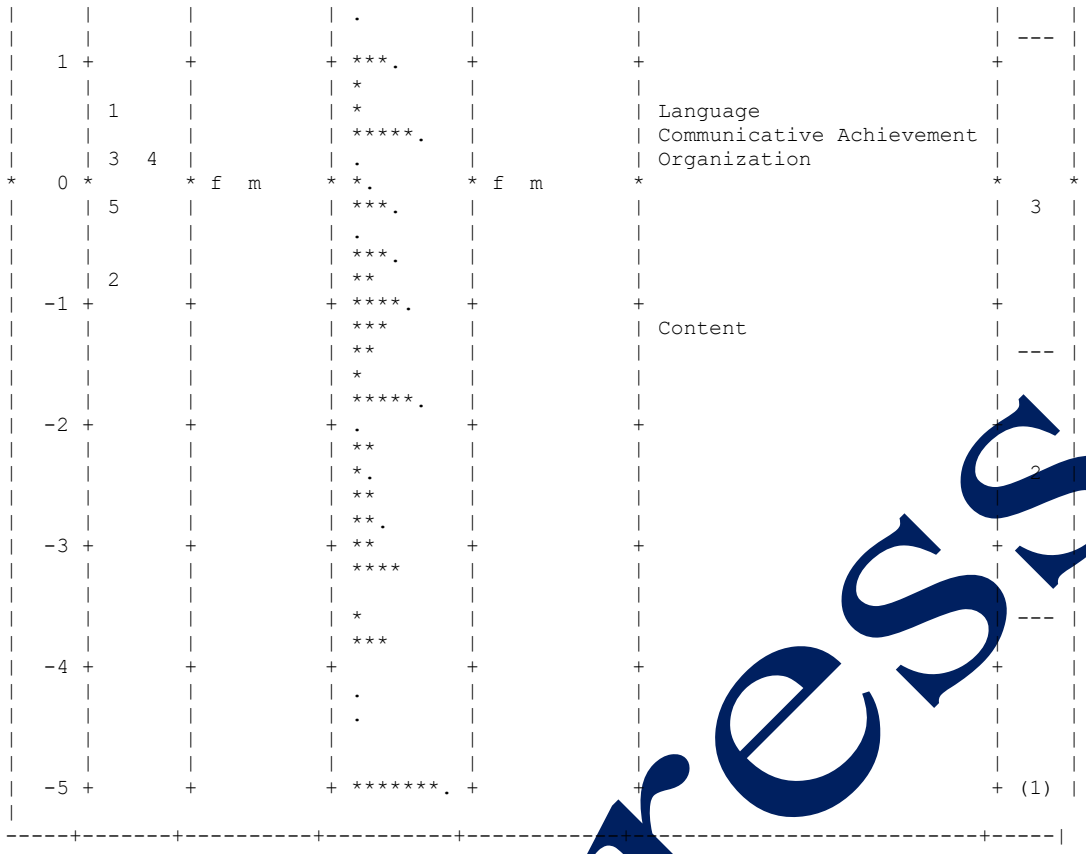
Rater Measurement Report

Rater	Total Score	Measure (SE)	Infit MNSQ	Outfit MNSQ
2	465	-.78	.91	.92
5	406	-.12	1.15	1.22
3	452	.14	.90	.91
4	405	.23	.75	.78
1	345	.53	1.34	1.33
Mean (SD)	414.6 (47.3)	0.00 (.49)	1.01 (.21)	1.03 (.21)

Figure 1

Variable Map of the Writing Performance Data





As Figure 1 and Table 1 indicate rater 1 is the most severe rater with a measure of .52 and rater 2 is the most lenient rater with a measure of -.78. The range of the rater measures is around 1.30 logits with a standard deviation of .49 which shows a high level of agreement among the raters.

Rater fit refers to the extent to which a given rater has assigned unexpected ratings. Infit and outfit values have an expected value of 1 and can range from 0 to infinity (Linacre, 2002; Myford & Wolfe, 2003). Raters with fit values greater than 1 tend to misfit the model. They show more variation than expected in their ratings. In contrast, raters with fit values less than 1 overfit the model and show less variation than expected in their ratings. Fit statistics smaller than 1 mean that the rater had muted ratings which suggests a central tendency or a halo effect (Engelhard, 2002; Myford & Wolfe, 2004). Linacre (2002) suggested the range 0.50 to 1.50 as an acceptable range for infit and outfit mean-square statistics. Other researchers suggested a narrower range of 0.70 to 1.30 and, for rating scales, a range of .60 to 1.40 (see, e.g., Bond et al., 2020; McNamara, 1996; Wright & Linacre, 1994). Our findings showed that rater infit and outfit measure are all within the acceptable range of .60 to 1.40 (ref) which shows that no rater behaved unexpectedly or had muted ratings.

3.1.2 Students' Measures

In this study, students Rasch ability estimates ranged from a low of -6.88 to a high of 6.88 with a mean and standard deviation of -1.10 and .85, respectively (including the examinees with minimum and maximum scores). The reliability of the measure was .87, with a separation of 2.59, and a strata value of 3.79. The mean of raw scores was 13.3 with a standard deviation of 7.

MFRM transforms ordinal raw scores of the students assigned by the raters into interval Rasch measures in logits. In this linear transformation, the severity of the raters is also considered, and ability estimates are adjusted for the severity of the raters.

Table 2.
Illustrative Results Comparing the Raw Scores and Facets Measures

Student	Raw score	Facets Masure	SE	Rater
128	17	-2.16	.53	4
68	17	-2.53	.53	3
55	19	4.20	1.14	2
156	19	4.58	1.13	5

Table 2 shows the raw scores and multifaceted Rasch measures of four students. Students 128 and 68 have the same raw scores of 17 but Facets has assigned them different ability estimates in logits, i.e., -2.16 and -2.53, respectively. Student 128 has been rated by rater 4 and student 68 has been rated by rater 3. According to Table 1, rater 4 is more severe than rater 3. Therefore, the ability estimates of student 128 is adjusted (raised) to compensate for the severity of the rater who has scored her essay. The same scenario can be observed for students 55 and 156. They both have the same raw scores but student 156 has a higher ability estimate based on MFRM. The reason is that rater 5, who has rated student 156's essay, is a harsher rater than rater 2 who has scored the essay of student 55.

3.1.3 Item Measures and Fit Statistics

Table 3 shows the item difficulty measures. Content is the easiest item and language is the hardest. In other words, raters have rated the content aspect of writing more generously while the language aspect is rated very harshly. The infit and outfit mean square values are all within the acceptable range which indicates unidimensionality and fit to the Rasch measurement model.

Table 3.
Item Measures and Fit Values

Item	Score	Measure	SE	Infit MNSQ	Outfit MNSQ
4	480	.54	.12	.95	1.00
2	493	.36	.12	.88	.98
3	504	.20	.12	1.04	1.08
1	596	-1.10	.12	1.05	1.05
Mean	518.3	0.00	.12	.98	1.01
(SD)	(52.8)	(.75)	(0.00)	(.08)	(.08)

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language, Pt-Measure Cor.=point-measure correlation

3.1.4 Rating Scale Diagnostics

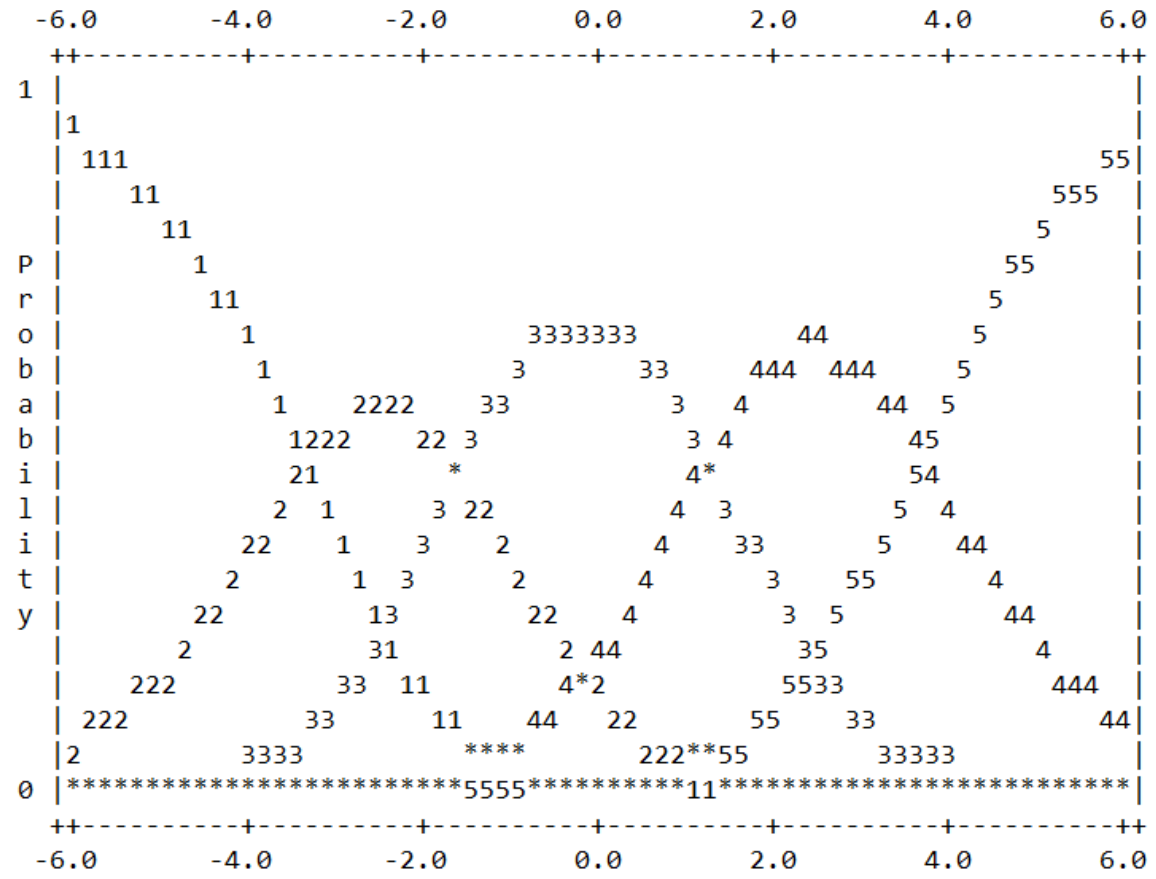
Table 4 displays the rating scale category statistics. The average measure is the average ability of people who respond in that category (Linacre, 2023). It is an empirical value that is computed and not an estimated Rasch model parameter. Average measures are expected to advance with categories and be ordered, which is the case here. Thresholds are the points on the rating scale where the probability of two adjacent categories is the same. It is a Rasch model parameter and is expected to be ordered. Disordered thresholds indicate an anomaly in the rating scale structure and interpretation (Andrich, 2013).

Table 4.
Rating Scale Statistics

Category	Category Counts	Average Measure	Threshold	Outfit MNSQ
1	152	-3.25	–	1.1
2	188	-1.98	-3.36	.9
3	273	-.45	-1.59	1.2
4	129	1.51	1.25	.9
5	42	3.60	3.70	.8

Figure 2 shows the probability of each category being assigned by the raters for different locations on the latent trait. Category 1 (score 1) is the most probable score for those examinees at the lower end of the ability scale. As ability increases the probability of other categories increases. The points where categories intersect are the threshold parameters. We expect a set of neat hills with peaks for a certain range of the ability scale (Bond et al., 2020).

Figure 2
Probability Curves for the Rating Scale



3.1.5 Interaction between Raters' Sex and Students' Sex

MFRM can reveal bias or interaction between different facets entered into the analysis. Table 5 reveals the interaction between the raters' sex and the examinees' sex. The table shows that male raters, on average, have given a measure of -.28 to the male students while female raters have given a measure of .22. That is, female raters have scored male examinees more leniently. The difference between these two values (i.e. contrast) is -.50. When this contrast is tested for statistical significance with a t-test, a t-value of -1.69, *df* =116 is obtained which is not significant, *p*=.09. That is, male and female raters have rated male examinees consistently and with no bias. The second row of Table 5 shows that male raters, on average, have given female students a measure of .06, while female raters have given them a measure of -.03. The contrast is very small (.09) which is not statistically significant *t* (460) =.60, *p*=.55.

Table 5.
Interaction between Rater Sex and Student Sex

Student Sex	Measure	SE	Rater Sex	Measure	SE	Rater Sex	Contrast	*t(df)*	*p*
M	-.28	.22	M	.22	.20	F	-.50	-1.69 (116)	.09

F	.06	.11	M	-.03	.09	F	.09	.60 (460)	.55

3.1.6 Interaction between Rater and Student Sex

Table 6 shows the bias or interactions between each of the five raters and the examinees' sex. Rater 1, on average has given a measure of .24 to male students while s/he has given a measure of .60 to female examinees. That is, s/he has rated females higher than males. The difference or contrast (.24-.60 = -.36) is not statistically significant t (49) =-.94, p =.35, i.e., this rater has rated males and females fairly. Table 6 shows that all raters have scored both genders fairly. Only rater 3 has rated male students, on average, .69 logit higher than female students. Although the contrast is not significant, it is very close to significance.

Table 6

Interaction between Raters and Student Sex

Rater	Measure	SE	Student Sex	Measure	SF	Student Sex	Contrast	*t(df)*	*p*
1	.24	.35	M	.60	.18	F	-.36	-.94(49)	.35
2	-1.05	.29	M	-.74	.15	F	-.31	-.96(66)	.34
5	-.27	.36	M	-.05	.17	F	-.22	-.54(39)	.58
4	.37	.34	M	.18	.17	F	.19	.51(40)	.61
3	.75	.33	M	.06	.14	F	.69	1.93(45)	.05

3.1.7 Interaction between Student Sex and Items

Interaction between students' sex and items is the familiar concept of differential item functioning or DIF. This analysis shows whether male and female examinees with equal ability estimates have the same chances of scoring on the items. Table 7 shows that none of the items exhibits DIF. According to the table, the difficulty of item 1 for males is .05 while it is .53 for females. That is, the item is harder for female students. The difference in the difficulty measures (contrast) is -.48 which is not statistically significant, t (60) = -1.48, p = .14.

Table 7

Interaction between Student Sex and Items

Item	Measure	SE	Student Sex	Measure	SE	Student Sex	Contrast	*t(df)*	*p*
2	.05	.29	M	.53	.14	F	-.48	-1.48(60)	.14
1	-1.09	.30	M	-1.15	.14	F	.06	.17 (60)	.86
4	.74	.29	M	.56	.14	F	.18	.57 (60)	.57
3	.31	.29	M	.07	.14	F	.24	.74 (60)	.46

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language

3.1.8 Interaction between Rater Sex and Items

Table 8 shows that male and female raters have given almost equal difficulty measures to the items. Male raters have given a difficulty measure of -.06 to item 3 while female raters have given it a measure of .36. That is, female raters have rated this item harsher. The contrast is -.42 which is not statistically significant t (151) = -1.69, p = .09.

Table 8

Interaction between Rater Sex and Items

Item	Measure	SE	Rater Sex	Measure	SE	Rater Sex	Contrast	t(df)	p.
3	-.06	.20	M	.36	.15	F	-.42	-1.69 (151)	.09

1	-1.21	.20	M	-1.05	.15	F	-.16	-.64 (152)	.52
4	.69	.20	M	.47	.15	F	.23	.91 (150)	.36
2	.57	.20	M	.25	.15	F	.33	1.31 (151)	.19

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language

3.1.9 Interaction between Raters and Items

Table 9 shows the bias or interaction between some items and raters. A few of these interactions are significant. The table shows the pairwise comparison of raters on items. For example, the first row of the table shows that rater 3 has given item 1 (content) a difficulty measure of -1.71 while rater 5 has given the item a difficulty of -.10. That is, rater 5 has rated this item more harshly that rater 3. The difference (contrast) is -1.61 which is statistically significant t (71) = -4.28, p = .0001.

Table 9

Interaction between Raters and Items

Item	Measure	SE	Rater	Measure	SE	Rater	Contrast	t(df)	*p*
1	-1.71	.26	3	-.10	.27	5	-1.61	-4.28 (71)	.0001
1	-1.55	.27	2	-.10	.27	5	-1.46	-3.83 (71)	.0003
3	-.56	.31	1	.84	.25	4	-1.40	-3.52 (60)	.0008
1	-1.25	.26	4	-.10	.27	5	-1.16	-3.11 (71)	.0027
3	-.56	.31	1	.28	.26	2	-.84	-2.09 (60)	.0413
3	-.56	.31	1	.27	.25	3	-.83	-2.07 (60)	.0428
1	-.75	.31	1	-.10	.27	5	-.65	-1.60 (59)	.1149
2	.22	.32	1	.81	.26	2	-.59	-1.44 (60)	.1560
3	.27	.25	3	.84	.25	4	-.57	-1.61 (75)	.1119
3	.28	.26	2	.84	.25	4	-.56	-1.57 (75)	.1209

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language

3.2 Rasch Rating Scale Analysis

The Rasch rating scale model (RSM, Andrich, 1978) was used to analyze the ratings given to the examinees. The RSM is a standard 2-facet model and, unlike the MFRM, does not accommodate raters and other aspects of the measurement. In the RSM analysis, the interaction between each rater and each rating criterion was considered an item. Because some examinees (n=40) had been rated by two raters, each rating criterion, i.e., content, communicative achievement, organization, and language, became two items. In other words, eight items were modeled altogether. However, since the value of the second version of the items was based on only 40 ratings and had a lot higher standard errors than their first version, the values of the first versions are reported and considered as the 'true' values of item difficulties here. For those who had been rated by one rater, the second version of the items was coded as missing. All the ratings were used for person estimation, though.

3.2.1 Item Measures and Fit Statistics

Table 10 shows the item measures and fit values for the four items of the test. As the table shows, Item 4, i.e., language is the hardest item, and content is the easiest. The infit and outfit mean square values are in the acceptable range which shows that they all fit the RSM and form a unidimensional construct of wiring ability in a foreign language. The point-measure correlation is the correlation between the item and the total score and is an index of item discrimination. Point-measure correlations should be positive and high.

Table 10

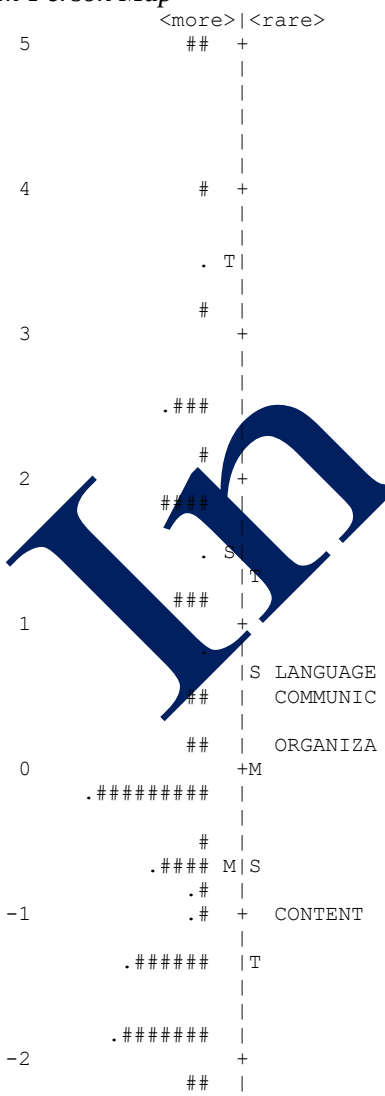
Item Measures and Fit Values

Item	Measure	SE	Infit MNSQ	Outfit MNSQ	Pt-Measure Cor.
4	.59	.13	.97	1.02	.82
2	.46	.13	.89	.93	.85
3	.18	.13	1.07	1.07	.85
1	-1.00	.14	1.15	1.22	.84

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language, Pt-Measure Cor.=point-measure correlation

Figure 3

Item-Person Map



```
                ####  |
                   .  |
                   # S|
                      |
   -3          .####  +
                      |
                   .  |
                      |
                ####  |
                      |
   -4                 +
                      |
                      |
                      |
                  .#  |
                     T|
   -5        #######  +
                  <less>|<frequ>
EACH "#" IS 2. EACH "." IS 1.
```

Figure 3 is the item-person map or the Wright map of the measurement. Items on the top of the scale are harder, and persons on the top are more proficient. The difficulty order of the items matches with the information presented in Table 9.

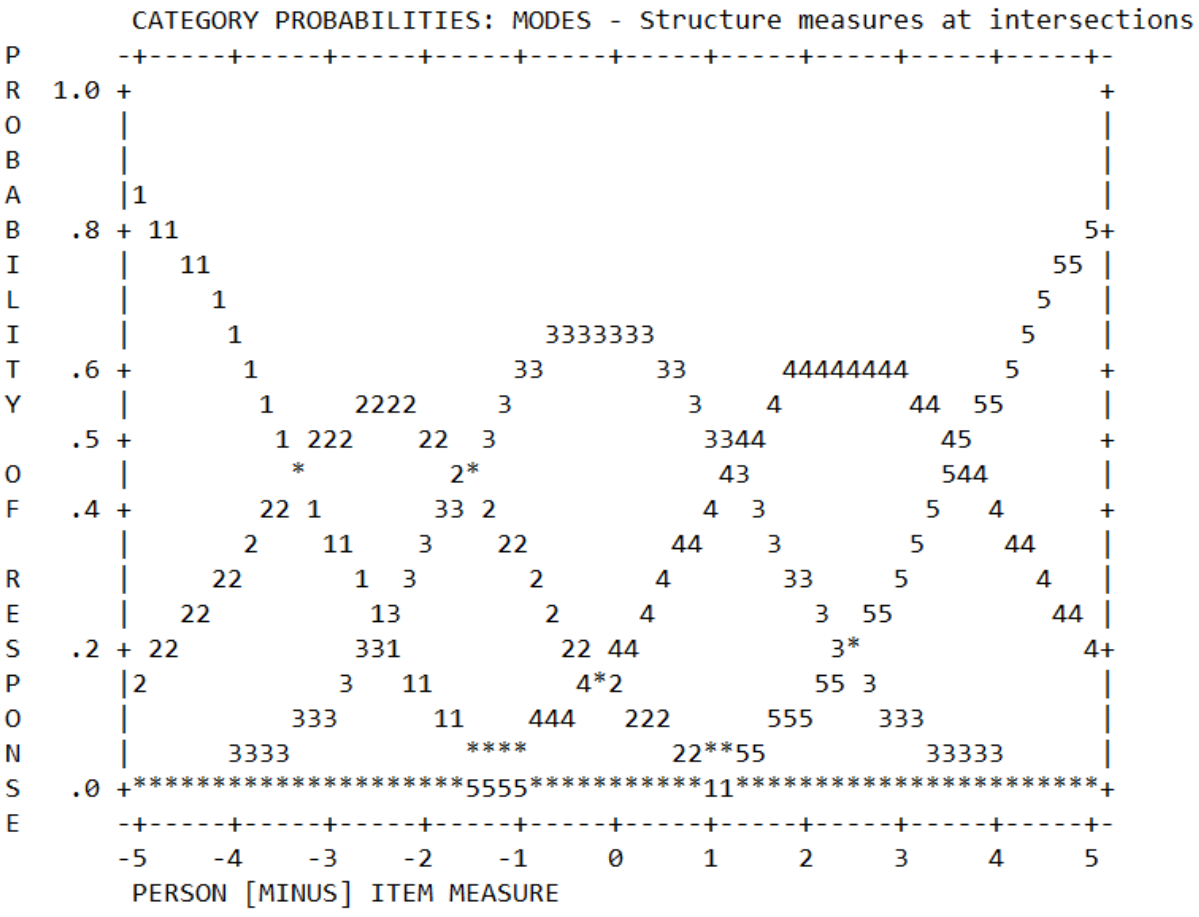
3.2.2 Rating Scale Diagnostics

Table 11 shows the rating scale statistics for the 5-point rating scale used to assign the ratings to the examinees. The interpretation of the values is the same as those in Table 4 for the MFRM. Figure 4 also gives the same information as Figure 2 above.

Table 11
Summary of Category Structure

Category	Category Counts	Observed Average	Threshold	Infit MNSQ	Outfit MNSQ
1	152	-3.11	_	1.14	1.17
2	188	-1.98	-3.30	.84	.86
3	273	-.45	-1.58	1.15	1.23
4	129	1.46	1.23	.85	.92
5	42	3.53	3.64	.80	.81

Figure 4
Probability Curves for the Rating Scale



3.2.3 Students' Measures

Table 12 presents the raw scores and RSM measures of the same four examinees who were rated by four different raters with varying levels of severity. As the table shows, RSM does not adjust ability estimates for the severity of the raters, and examinees with the same raw scores have the same Rasch ability estimates. The RSM-based person measures are in Appendix 3.

Table 12

Illustrative Results Comparing the Raw Scores and RSM Measures

Student	Raw score	RSM Masure	SE	Rater
128	17	-2.19	.52	4
68	17	-2.19	.52	3
55	19	4.97	1.13	2
156	19	4.97	1.13	5

3.2.4 Differential Item Functioning

Interaction between students' sex and items is referred to as differential item functioning (DIF) in the standard 2-facet measurement. Table 13 shows the difficulty (measure) of each item among girls (Class F) and boys (Class M). The difference between the difficulties in these two classes (Contrast) is tested for statistical significance using a t-test. The table shows that Item 1 has a difficulty of -1.04 for girls and a difficulty of -1.12 for boys. That is the item is slightly easier for boys. However, the difference (.08) is not statistically significant $t(45)=.23$, $p=.81$. In other words, the items are equally easy or hard for girls and boys. Or raters have rated girls and boys with the same level of harshness or leniency. Item 2 has a difficulty of .72 for girls and a difficulty of -.22 for boys. That is girls have found this item harder. The difference in difficulty is .94 which is statistically significant $t(45)=2.54$, $p=.01$.

The other interpretation is that raters have rated girls harsher on this item than the boys. In other words, there is an interaction between raters' and examinees' sex.

Table 13
Differential Item Functioning by Examinees' Sex

Class	Measure	SE	Class	Measure	SE	Contrast	*t*(*df*)	*p*	Item
F	-1.04	.16	M	-1.12	.34	.08	.23 (45)	.81	1
F	.72	.16	M	-.22	.33	.94	2.54 (45)	.01	2
F	.06	.16	M	.11	.33	-.05	-.14 (45)	.88	3
F	.64	.16	M	.56	.33	.09	.23 (45)	.81	4

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language

3.2.5 Interaction between Raters' Sex and Items

Differential item functioning was run by raters' sex. The classes in Table 14 refer to the raters' sex. Item 1 is given a difficulty measure of -.89 by female raters while male raters have given it a difficulty of -1.17. That is, female raters have rated 'content' more severely than male raters. However, the difference in the difficulties (.27) is not statistically significant, t (127) =.99, p=.32. Table 13 shows that these differences are not significant for any of the items. In other words, there is no interaction between raters' sex and items.

Table 14
Differential Item Functioning by Raters' Sex

Class	Measure	SE	Class	Measure	SE	Contrast	*t*(*df*)	*p*	Item
F	-.89	.17	M	-1.17	.21	.27	.99 (127)	.32	1
F	.32	.17	M	.70	.22	-.38	-1.40 (126)	.16	2
F	.35	.17	M	-.08	.21	.43	1.56 (126)	.12	3
F	.46	.17	M	.80	.22	-.33	-1.21 (126)	.22	4

Note: Item 1= Content, Item 2= Communicative Achievement, Item 3= Organization, Item 4= Language

3.2.6 Interaction between Items and Raters

Table 15 is the truncated DIF table by the rater. It shows the pairwise comparison of item measures by raters. It is analogous to MFRM Table 9 above. The person classes here refer to the individual raters. The first row of the table shows that Item 1 (content) is given a difficulty of -.47 by rater 1 while it is given a difficulty of -1.87 by rater 2. That is, rater 1 has rated this item more severely than rater 2. The contrast is 1.40 which is statistically significant t (54)=3.20, p=.002. Table 14 shows that several other rater-by-item interactions are statistically significant.

Table 15
Differential Item Functioning by Rater

Class	Measure	Class	Measure	Contrast	*t*(*df*)	*p*	Item
1	-.47	2	-1.87	1.40	3.20 (54)	.0023	1
1	-.47	3	-1.91	1.43	3.28 (54)	.0018	1
1	-.47	4	-1.02	.55	1.30 (54)	.1987	1
1	-.47	5	.33	-.80	-1.84 (51)	.0715	1
1	.49	2	.89	-.40	-.92 (54)	.3605	2
1	.49	3	.74	-.25	-.59 (54)	.5559	2
1	.49	4	-.02	.51	1.21 (54)	.2322	2
1	.49	5	.23	.26	.58 (51)	.5622	2
1	-.28	2	.11	-.39	-.91 (54)	.3660	3
1	-.28	3	.41	-.69	-1.62 (54)	.1101	3

3.2.7 Dimensionality

Principal components analysis of standardized residuals as a technique to examine unidimensionality (Linacre, 2023) showed that the strength of the first contrast extracted from the residuals is 2.3 which is greater than the minimum value of 2 suggested by Linacre (2023). This is evidence that the writing data evaluated in this study is not unidimensional. This is in line with Effatpanah and Baghaei (2024) who came to the same conclusion in the context of Kim's checklist for writing evaluation.

4. Discussion

In this section, the statistics that are provided by the two models (MFRM and RSM) and their associated software (Facets and Winsteps) are compared and discussed. The comparisons are made based on the corresponding tables from the two analyses.

4.1 Rater Severity

Table 1 shows the severity and fit statistics of the raters. Since RSM is a 2-facet model and only accommodates items and persons, no statistics for individual raters are produced by RSM. In the RSM, the difficulty of the items and the severity of the raters are conflated and cannot be distinguished. The advantage of the MFRM is that we can separate task or item difficulty from rater severity. Comparing Figure 1 and Figure 4 shows that the MFRM parameterizes all the facets of the measurement on a common scale but the RSM, being a 2-facet model, only locates persons and items on a common scale. RSM does not give any fit statistics for the raters either. That is, with RSM we cannot evaluate whether the raters rated the examinees consistently, unexpectedly, or tended to use the center of the sale.

4.2 Students' Measures

Table 2 and Table 12 show that while MFRM adjusts examinees' scores for raters' severity, the RSM does not consider the impact of raters' harshness on students' ability estimates. Table 12 shows that examinees with the same raw scores have the same Rasch ability estimates regardless of the severity of the rater who scored their essays. Table 2, in contrast, clearly shows that when an examinee's essay

is rated by a harsh rater, her ability estimates is augmented proportionately to compensate for the rater's harshness. Further analyses showed that the correlation between the MFRM-based person measures and RSM-based measures was .989 which suggests near-perfect consistency between the ability estimates yielded by the two models. However, when the absolute values of the differences between the estimated person measures (from the two models) for each student are examined, a different picture emerges. These absolute differences ranged from zero to .90 with a mean of .30 and a standard deviation of .24. That is, a person's writing ability estimates can change by .90 logit depending on the selected model. On average an examinee's ability estimates could be different by .30 logit depending on the model. This could be a huge difference, especially around the cut-points where pass/fail decisions are made.

4.3 Item Measures and Fit Statistics

Table 3 and 10 show the item difficulty measures and their fit values in the two models. Comparison of the two tables shows that the item measures are very close and quite comparable. The standard errors of the difficulty parameters and the fit values are also very close. Item 1 (content) has a poorer fit in the RSM while its fit is very good in the MFRM. Note that, the differences in the estimates and fit values are because in the RSM each item was estimated twice, i.e., eight items were estimated. The deviance statistic (-2loglikelihood) of the two models showed that the MFRM with a deviance of 1320.42 fits better than the RSM with a deviance of 1339.70. Comparison of the item infit and outfit statistics across the two models showed that although the items have acceptable values in both models, they are closer to perfect fit in the MFRM. That is, the MFRM fits the data better than the RSM both at the global and item level.

4.4 Rating Scale Diagnostics

Table 4 and Table 11 show the rating scale statistics based on MFRM and RSM, respectively. Rating scale statistics from both models show that the thresholds are ordered and average measures advance with category scores. Comparison of the two tables shows that the values of the threshold parameters and other statistics from the two models are very close. Comparison of Figure 2 and Figure 3 also shows that the thresholds are ordered and the 5-point category scale works as expected.

4.5 Interaction between Raters' Sex and Students' Sex

MFRM allows researchers to examine the interaction between all facets. Table 5 shows the interaction between raters' sex and students' sex. RSM does not provide this diagnostic. In fact, interactions within a 2-facet model can be examined in the framework of DIF. That is, with a 2-facet model, bias and interaction can be evaluated only when items are involved. Since in the interaction of raters' sex and examinees' sex items are not involved, the RSM does not yield any information concerning this bias.

4.6 Interaction between Raters and Student Sex

Table 6 shows the interaction between raters' sex and students. It reports whether individual raters have systematically favored or disfavoured a particular sex. This piece of information is only provided by MFRM and RSM does not provide this.

4.7 Interaction between Student Sex and Items

Tables 7 from MFRM and Table 13 from RSM are analogous. The two tables provide the same information about the performance of the male and female students on the four items or rating criteria. Table 7 shows that there is no interaction between the items and the examinees' sex. That is, both male and female students have found the items equally difficult. However, Table 12 shows that Item 2 (communicative achievement) exhibits DIF. This item is .94 logit easier for male students which is statistically significant. Table 7 shows that this item is .48 logit easier for male students which is non-significant.

4.8 Interaction between Rater Sex and Items

Table 8 from MFRM and Table 14 from RSM are analogous. Table 8 shows that male and female raters have rated all four items with equal severity. Table 13 also confirms this finding.

4.9 Interaction between Raters and Items

Table 9 and Table 15 are analogous. They show the pairwise comparison of raters on individual items. This information is provided by both MFRM and RSM. For illustration purposes, we compare row 1 of Table 9 with its corresponding row in Table 14 (the shaded row). These rows compare the ratings of raters 3 and 5 on content. According to Table 9, rater 3 has given content a difficulty of -1.71 while rater 5 has given it a difficulty of -.10. That is, rater 5 has rated this item very harshly. The difference in these two measures is (contrast) -1.61 which is statistically significant. Based on Table 14, rater 3 has given content a measure of -1.91 but rater 5 has given it a measure of .33, a contrast of -2.24 which is statistically significant. That is, the two models confirm that raters 3 and 5 have rated 'content' with different levels of harshness. The other pairwise comparisons can be matched in the two analyses using Tables 9 and 14.

4.10 Dimensionality

Principal components analysis of standardized residuals from RSM showed that the data are not unidimensional but the MFRM does not provide this information. This, however, is not a shortcoming of MFRM but a shortcoming of the Facets software program which does not contain this analysis.

5. Conclusion

This study compared the Multifaceted Rasch Model (MFRM) and the Rasch Rating Scale Model (RSM) for evaluating English as a Foreign Language (EFL) writing performance. The results provide insights into the effectiveness of each model in handling rater-mediated assessments and highlight key differences in their ability to account for rater variability, fairness, and diagnostic insights.

The study showed that the MFRM analysis disclosed significant variation in rater severity, with a range of 1.30 logits and a standard deviation of 0.49. This suggests that raters differ in their scoring tendencies, which aligns with previous research on rater bias in performance-based assessments (Eckes, 2015; Engelhard, 2013). However, all raters demonstrated acceptable fit values within the standard range (0.60–1.40), indicating consistent and reliable scoring patterns.

One of the most significant findings was that MFRM adjusts students' ability estimates based on rater severity, whereas RSM does not. This was evident in cases where two students received identical raw scores but were rated by different raters with varying severity levels. The MFRM-adjusted measures accounted for this discrepancy, ensuring fairer evaluations, while RSM produced identical ability estimates for students with the same raw scores, disregarding rater effects. This suggests that MFRM provides a more nuanced and equitable assessment of writing performance.

Both models showed similar item difficulty rankings, with content being the easiest and language being the most difficult. This pattern aligns with expectations, as linguistic accuracy is often assessed more strictly than content (McNamara, 1996). Additionally, both models showed acceptable item fit values, confirming the validity of the rating criteria. However, the MFRM fit statistics were slightly better, indicating a superior model fit.

MFRM revealed several important interaction effects. There was an interaction between rater gender and student gender, but there were no significant biases, although female raters tended to score male students more leniently. There was an interaction between rates and students While most raters scored both genders fairly, one rater (Rater 3) exhibited near-significant bias toward male students. RSM detected one DIF item (communicative achievement), which was significantly harder for female students. MFRM, however, did not identify this as significant.

The principal components analysis (PCA) of standardized residuals from RSM showed that the data were not unidimensional, aligning with previous research (Effatpanah & Baghaei, 2024). However, MFRM did not provide this diagnostic, which is a limitation of the FACETS software rather than the model itself.

Overall, our findings indicate that MFRM provides more precise, fairer assessments by adjusting for rater severity and bias. This makes it particularly useful for high-stakes tests like IELTS or TOEFL, where rater variability can impact student outcomes. In criterion-referenced tests where decisions are based on the scores and pass/fail decisions are made, it is crucial to use the MFRM. If the decision-making process is norm-referenced, either model can be utilized effectively. In criterion-

referenced contexts, the choice of the model can significantly impact outcomes and lead to notable differences.

The study supports the use of MFRM for rater-invariant measurement but also highlights its computational complexity and limited accessibility. The lack of user-friendly MFRM software may hinder its widespread adoption. While RSM is easier to implement, its failure to account for rater effects may lead to less reliable evaluations. Institutions should consider adopting MFRM for more equitable assessments, particularly in performance-based testing.

The findings suggest that MFRM outperforms RSM in evaluating EFL writing performance due to its ability to adjust for rater severity, detect interactions, and ensure fairer scoring. However, RSM remains a viable alternative in contexts where computational simplicity is prioritized. Future research should explore ways to enhance MFRM accessibility, such as developing more user-friendly software and training raters and educators on its implementation.

To ensure the effective use of MFRM in rater-mediated assessments, structured training programs should be developed for educators, raters, and assessment designers. These programs should include Introductory Workshops covering the fundamentals of Rasch measurement, MFRM concepts, and the importance of rater-invariant scoring. Hands-on sessions using FACETS, Winsteps, or other Rasch-based software to help raters interpret MFRM outputs and apply adjustments. Allowing educators to analyze real-world assessment data, identify rater biases, and compare MFRM-based adjustments with raw scores. Institutions can establish certification courses to ensure that raters are well-equipped to apply MFRM in their assessment frameworks. By integrating interactive training modules and continuous professional development, institutions can empower educators to apply MFRM more effectively, ensuring fairer and more accurate assessments.

For institutions to successfully incorporate MFRM into large-scale assessments, a systematic adoption strategy is necessary. Standardizing MFRM adoption in assessment policies to ensure rater-invariant measurement practices across various testing programs. Embedding MFRM-compatible software into assessment platforms used for grading, such as AI-enhanced grading systems or LMS-integrated Rasch analysis tools. Establishing centralized databases where rater performance metrics (e.g., severity indices, bias detection) are continuously monitored and analyzed. Before full-scale implementation, institutions should conduct pilot programs to refine MFRM scoring frameworks and validate their effectiveness against traditional scoring models. Implementing continuous rater calibration sessions, where MFRM-generated reports are regularly reviewed to ensure consistency and fairness in scoring. By following this framework, institutions can seamlessly integrate MFRM into large-scale assessment systems, ultimately leading to fairer, data-driven evaluations in high-stakes testing environments.

While the statistical results indicate that MFRM provides a better model fit than RSM, it is important to clarify why this fit is significant in real-world assessment settings. In practical terms, a better model fit translates into More Accurate Student Ability Estimates. Since MFRM accounts for rater severity, it ensures that students are evaluated fairly, independent of who scores their work. This is particularly critical in high-stakes exams (e.g., TOEFL, IELTS), where small scoring inconsistencies could impact admission decisions. Institutions can use MFRM outputs to detect and correct rater biases in real time, improving the reliability of scoring in large-scale assessments. A well-fitting model ensures that scores remain consistent over time, enabling fairer comparisons between different test administrations. When scores are used for placement, certification, or policy decisions, a model with a better fit reduces the likelihood of unfair classifications, thereby improving the validity of the assessment. By demonstrating these advantages, this study highlights why MFRM is not just a superior statistical model but also a more practical and equitable tool for real-world educational assessments.

Beyond Rasch-based approaches, there are several other assessment models that institutions might consider. This section briefly compares MFRM, Classical Test Theory (CTT), and Generalizability Theory (G-Theory) in the context of rater-mediated assessments. MFRM explicitly adjusts for individual rater severity and leniency, ensuring that student scores are independent of who

evaluates their work. In contrast, CTT assumes that rater effects are random and does not account for them, leading to potential inconsistencies. G-Theory provides a more sophisticated approach by analyzing multiple sources of error, including rater effects, but does not always adjust for bias at the individual level.

MFRM is suitable for both small- and large-scale assessments, making it an adaptable option for various testing environments. CTT works best with large samples, but it is less effective when scoring variability arises from rater differences. G-Theory is highly scalable and works well in large-scale assessments, particularly in educational research settings where multiple sources of variance need to be analyzed. One of the biggest strengths of MFRM is that it provides rater-invariant scoring, ensuring fairness by adjusting scores for differences in rater severity. CTT does not account for rater variance, meaning that scores may fluctuate depending on test difficulty and rater tendencies. G-Theory allows researchers to identify and estimate variance components, but it does not inherently adjust for bias, making it less useful in practical rater-mediated assessment settings.

MFRM produces logit-based ability scores that are adjusted for rater bias, providing a more refined measurement of student ability. CTT relies on raw scores, which are easier to understand but are more vulnerable to rater subjectivity. G-Theory generates variance estimates for different sources of error (e.g., rater bias, task difficulty), but it requires complex statistical modeling, making it less accessible for general practitioners. MFRM is particularly well-suited for high-stakes testing environments, such as TOEFL, IELTS, and teacher evaluations, where rater bias must be controlled. CTT remains widely used in traditional standardized tests (e.g., SAT, ACT), where rater effects are less relevant. G-Theory is valuable in large-scale educational research and multi-faceted assessments, but its complexity makes it more challenging to implement in routine educational assessments.

In summary, MFRM is the best option for rater-mediated assessments because it explicitly models and adjusts for rater bias, ensuring fairness and accuracy. CTT is widely used but does not separate rater effects from other sources of error, making it less reliable in scoring subjective performance tasks. And finally, G-Theory is powerful for estimating multiple sources of error, but it requires more statistical expertise and does not always provide direct bias adjustments like MFRM.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Athuors.(under review). The application of Rasch rating scale model in evaluating writing performance.

Cambridge English (2024). *Assessing writing for Cambridge English Qualifications: A guide for teachers*. Cambridge University Press & Assessment. https://www.cambridgeenglish.org/images/600975-teacher-guide-for-writing-b2-first-for-schools.pdf

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

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

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

Effatpanah, F., & Baghaei, P. (2024). Examining the dimensionality of linguistic features in L2 writing using the Rasch measurement model. *Educational Methods & Practice*, *2*, 1–22.

Engelhard, G. (2002). Monitoring raters in performance assessments. *In G. Tindal & T. Haladyna (Eds.), Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Lawrence Erlbaum Associates.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. *International Handbook of English Language Teaching*, *1*, 487–504. https://doi.org/10.1007/978-0-387-46301-8_35

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating scales for TOEFL iBT integrated writing tasks*. TOEFL iBT Research Report, RR-13-05. Educational Testing Service.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). American Council on Education/Praeger.

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

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

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

Linacre, J. M. (2014). *Facets computer program for many-facet Rasch measurement*. Winsteps.com.

Linacre, J. M. (2023). *A user's guide to WINSTEPS and Facets*. Winsteps.com.

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

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532212452208

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.

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

Shahi, R. , Ravand, H. & Rohani, G. R. (2025). Examining the effect of item difficulty and rater leniency on Iranian test Takers' performance on WDCT and DSAT: A comparative study. *International Journal of Language Testing, 15*(1), 1–19. doi: 10.22034/ijlt.2024.454478.1341

Taufiqulloh, T. , Fadhly, F. Zaman , Rosdiana, I. , Ferrer, C. N , Ratsamemonthon, C. , Nindya, M. A. & Irawan, N. (2025). Comprehensive review of writing assessments in EFL contexts: A meta-synthetic study. *International Journal of Language Testing, 15*(1), 193–213. doi: 10.22034/ijlt.2024.475218.1367

Wind, S. A., & Peterson, M. H. (2018). A systematic review of the validity of the Rasch model in language assessment research. *Educational and Psychological Measurement, 78*(4), 589–610. https://doi.org/10.1177/0013164416671039

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

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.