SCOPE
Journal of English Language Teaching
| p-ISSN 2541-0326 | e-ISSN 2541-0334 |
https://journal.lppmunindra.ac.id/index.php/SCOPE/

Research Article

Rater Severity/Leniency and Bias in EFL Students' Composition Using Many-Facet Rasch Measurement (MFRM)

Yenni Arif Rahman[1], *Fitri Apriyanti*[2] , *Rahmi Aulia Nurdini*[3]

[123] *English Department, Faculty of Language and Communication, Universitas Bina Sarana Informatika*
Jl. Kramat Raya No.98, RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450, Indonesia

KEYWORDS

Rater Severity/Leniency;
Rater Bias;
EFL Composition;
Many-Facet Rasch Measurement.

CORRESPONDING AUTHOR(S):

E-mail: yeni.yar@bsi.ac.id*

A B S T R A C T

The study aims to investigate the extent to which raters exhibit tendencies towards being overly severe, lenient, or even bias when evaluating students' writing compositions in Indonesia. Data were collected from 15 student essays and four raters with master's degrees in English education. The Many-facet Rasch measurement (MFRM), automatized by Minifac software, a program created for the Many-facet Rasch measurement, was used for data analysis. This was done by meticulously dissecting the assessment process into its distinct components—raters, essay items, and the specific traits or criteria being evaluated in the writing rubric. Each rater's level of severity or leniency, essentially how strict or lenient they are in assigning scores, is scrutinized. Likewise, the potential biases that raters might introduce into the grading process are carefully examined. The findings revealed that, while the raters used the rubric consistently when scoring all test takers, they varied in how lenient or severe they were. Scores of 70 were given more frequently than the other score. Based on the findings, composition raters may differ in how they rate students which potentially leading to student dissatisfaction, particularly when raters adopt severe scoring. The bias in scoring has highlighted that certain raters consistently tend to inaccurately score items, deviating from the established criteria (traits). Furthermore, the study also found that having more than four items/criteria (content, diction, structure, and mechanic) is essential to achieve a more diverse distribution of item difficulty and effectively measure students' writing abilities. These results are valuable for writing departments to improve the oversight of inter-rater reliability and rating consistency. To address this issue, implementing rater training is suggested as the most feasible method to ensure more dependable and consistent evaluations.

INTRODUCTION

In the Indonesian educational landscape, students encounter notable challenges when it comes to developing proficient writing skills in English as a Foreign Language (EFL). Cahyono (2019) points out, writing is often considered the most difficult skill to acquire due to its complexity, involving the mastery of vocabulary, grammar, and rhetorical conventions. These difficulties

DOI: http://dx.doi.org/10.30998/scope.v8i1.19432

encompass issues such as grammar accuracy, vocabulary selection, sentence structure, and coherent organization of ideas. This situation is exacerbated by the linguistic differences between Bahasa Indonesia and English, leading to instances of literal translation and non-idiomatic language use (Rahayu & Widiati, 2019; Setiawan & Hartoyo, 2020).

Moreover, the urgency of assessing students' writing skills in Indonesia cannot be overstated. The lack of effective writing assessment strategies hinders educators' ability to gauge students' progress accurately and provide targeted instruction to address specific weaknesses. Linacre (2006) emphasizes the importance of proper assessment, stating that assessment not only measures but also guides instruction, offering insights into learners' strengths and areas needing improvement. Without comprehensive assessment tools, students' writing potential may remain untapped, hindering their ability to effectively communicate in written English.

In this context, the relationship between students' writing difficulties and the severity and leniency of raters becomes a crucial consideration. Rater biases can significantly impact the reliability and validity of writing assessments (Cahyono, 2019). Thus, a comprehensive understanding of raters' tendencies towards being excessively severe or lenient is imperative to ensure fair and equitable evaluations of students' writing proficiency. Rater severity/leniency in students' composition refers to the degree to which raters consistently assign higher or lower scores when evaluating the written work of students (Myford & Wolfe, 2004; Huang, 2023). It is an important aspect of assessment in various educational contexts, including language education and writing courses. Rater severity can have a significant impact on students' grades and overall assessment outcomes, as it directly influences the perceived quality of their compositions. Understanding rater severity is crucial for ensuring fair and accurate evaluations, as well as providing meaningful feedback to students to support their learning and improvement in writing skills (Li, 2022).

The presence of rater severity introduces challenges in achieving consistent and reliable assessment practices (Wind, 2019). When raters exhibit high severity, they tend to assign lower scores more frequently, potentially underestimating the true abilities of students. Conversely, when raters display leniency or low severity, they may assign higher scores more often, potentially inflating the perceived quality of the compositions (Ahmadi Shirazi, 2019; Erguvan & Aksu Dunya, 2020). Both scenarios can lead to discrepancies in grading and affect the fairness of the assessment process. Therefore, it becomes essential to analyze and address rater severity to enhance the validity and reliability of evaluation in students' composition.

Various factors can contribute to rater severity in students' composition. These factors include individual rater characteristics, such as personal biases, preferences, and experience levels (Fahim & Bijani, 2011; Huang, 2023; Noor, Beram, Huat, Gengatharan, & Mohamad Rasidi, 2023). Raters with strict personal standards may exhibit higher severity, while more lenient raters may display lower severity (Tanaka, 2023). Additionally, the characteristics of the writing prompts or tasks assigned to students can also influence rater severity. The complexity, clarity, and specific requirements of the prompts can impact raters' interpretations and subsequently affect their severity in assigning scores (Lim, 2009). Understanding these contributing factors is crucial for implementing effective strategies to minimize the impact of rater severity and improve the consistency and fairness of assessments in students' composition (Uto, 2022).

In recent years, there has been a growing interest in exploring the concept of rater severity in various educational and assessment contexts. One approach gaining momentum in this area is the application of the Rasch measurement model to examine rater severity in a novel and comprehensive way (Bond & Fox, 2007; Linacre, 2002; Maier, 2001). The Rasch measurement provides a robust framework for evaluating rater behavior by considering both the difficulty of the items being rated and the ability of the raters themselves (Bond & Fox, 2015; Sumintono & Widhiarso, 2013; Tan, 2013). By utilizing the Rasch measurement, researchers can investigate how raters differ in their level of severity, which has implications for ensuring fairness and consistency in scoring and evaluation processes. This innovative research approach not only enhances our understanding of rater behavior but also offers insights into ways to improve the validity and reliability of assessments.

One of the leading studies regarding rater severity was done by Myford & Wolfe (2004) who are ones of the pioneers to introduce the use of Multi-Facet Rasch Measurement for detecting and measuring raters' effects. They discuss the Facets (Linacre, 2020) computer program to study five of raters' effects: leniency or severity, central tendency, randomness, halo, and differential leniency/severity. This seminal work led other researchers to these fields. The recent research on rater severity was conducted by Erguvan and Aksu Dunya (2020) using MFRM to examine students' composition with multi-trait rubrics with ESL participants. The finding shows that composition instructors may differ in their rating behavior, and this may cause dissatisfaction, creating a sense of unfairness among the students of severe instructors. This research also demonstrates the importance of rater training to consolidate the raters scoring consensus.

While rater severity has been extensively investigated in various educational domains, limited attention has been given to its application within the EFL setting. Understanding the influence of rater characteristics in this context is crucial due to the unique challenges faced by EFL students, such as language proficiency limitations and cultural differences. By examining rater severity in the evaluation of EFL compositions, this research contributes to the development of tailored assessment practices that align with the specific needs and circumstances of these students. This approach acknowledges the distinctiveness of EFL writing and aims to promote more accurate and constructive feedback, ultimately fostering improved language learning outcomes for this population.

Furthermore, the application of the Rasch measurement in studying rater severity allows for the identification and quantification of various sources of rater bias or leniency. Researchers can examine how individual raters differ in their inclination to assign higher or lower scores and identify potential factors contributing to such biases. This approach enables a more nuanced analysis of rater behavior beyond mere average score differences, shedding light on the underlying patterns and tendencies that influence rating outcomes. By uncovering these sources of rater bias, educational institutions, assessment agencies, and researchers can develop targeted interventions and training programs to mitigate the impact of bias, improve rating accuracy, and ensure fair evaluations. Ultimately, this novelty research contributes to the advancement of assessment practices, leading to more reliable and valid results in various fields, including education, psychology, and performance evaluations.

METHOD

This study involved four raters with master degree in English education and 15 EFL students who had completed the TOEFL iBT essay writing course. The participants consisted of high school and university students, assumed to have an Intermediate or higher level of English proficiency, as evident from their TOEFL scores above 500. The essays were expected to follow a standard structure comprising an introduction, content, and conclusion, totaling five paragraphs. This uniform structure aimed to prevent bias in the raters' assessment, as the number of paragraphs could otherwise influence their judgment. Additionally, this standardized format served as a reference for determining whether a student's writing could be used as an essay sample, ensuring appropriate validation of research results. The essay samples were collected by the Academic Writing instructor, who explicitly provided writing instructions to the students.

The assessment of rater severity in this study used the Many Facets Rasch Measurement (MFRM) method. This method is employed to measure regression patterns among various raters assessing students' writing. The software used for this purpose is FACETS Minifac, developed by Linacre (2020).

There are four writing skills (items) to be evaluated: content, structure, diction, and mechanics. Each items has provided multiple traits belong to certain rating. The assessment criteria utilized by the raters are the holistic scoring rubrics developed by (Jacobs., Holly, Stephen, Zingkgraf, Deanne, Wormuth, Faye, Jane, 1981).

Table 1. Holistic Score: Rating and Criteria (Jacob et al., 1981)

Rating	Criteria
Proficient	1. Writes single or multiple paragraph with clear introduction, fully develop idea, and clear introduction 2. Uses appropriate verb tense and a variety of grammatical and syntactical structures; uses complex sentences effectively; uses smooth transitions 3. Uses varied, precise vocabulary 4. Has occasional errors in mechanics (spelling, punctuation, and capitalization) which do not detract from meaning
Fluent	1. Writes single or multiple paragraph with main idea and supporting detail, present idea logically, though some parts may not fully developed 2. Uses appropriate verb tense and a variety of grammatical and syntactical structures; errors in sentence do not detract from meaning; uses transitions 3. Uses varied vocabulary appropriate for the purpose 4. Has few errors in mechanics which do not detract from meaning
Expanding	1. Organizes ideas in logical or sequential order with some supporting detail; begins to write a paragraph 2. Experiment with a variety of verb tenses, but does not use them consistently; subject/verb agreement errors; uses some compound and complex sentences; limited use of transitions 3. Vocabulary is appropriate to purpose but sometimes awkward 4. Use punctuation, capitalization, and mostly conventional spelling; errors sometimes interfere with meaning
Developing	1. Writes sentences around an idea; some sequencing present, but may lack of cohesion 2. Write in present tense and simple sentences; has difficulty with subject/verb agreement, run-on sentences are common; begin to use compound sentences 3. Uses high frequency words; may have difficulty with word order; omit endings or words 4. Uses some capitalization, punctuation and transitional spelling; errors often interfere with meaning

 DOI: http://dx.doi.org/10.30998/scope.v8i1.19432

Beginning	1. Begin to convey meaning through writing 2. Write predominantly phrases and patterned or simple sentences 3. Uses limited or repetitious vocabulary 4. Uses temporary (phonetic) spelling
Emerging	1. No evidence of idea development or organization 2. Uses single word, pictures, and patterned phases 3. Copies from model 4. Little awareness of spelling, capitalization, or punctuation

In this research, scaling is compulsory to convert the regular score (example 70 or 80) so the score can be processed in Minifac software. The scaling is conducted using the Likert scale, which consists of five response options presented in Table 1: proficient, fluent, expanding, developing, beginning, and emerging (merged as one group). The rubric rating scale provided in Table 2 is then interpreted in the following manner: [details about the interpretation would be provided in Table 2.

Table 2. Rubric Rating Scale

Scale	*Likert* Score
Proficient	5
Fluent	4
Expanding	3
Developing	2
Emerging & Beginning	1

There are several types of data that need to be prepared for measuring rater severity. The first one is raw data, which includes the raters, student composition, items, and composition scores assigned by the raters. Next, this raw data is input into an Excel spreadsheet and converted into text format, which is then processed and entered into the prepared syntax. The adapted syntax coding for this research can be found in the appendix. At this point, the data is ready to be analyzed using the Minifac software (all details of data and syntax coding is attached in appendix).

The next step involves inputting the coded data from the notepad into the Minifac software. There are four output data from Minifac which will be interpreted: unidimensionality test to measure the construct/item validity, rater validity test by examining the Outfit Mean Square (MNSQ) results in the Rater Measurement Report, Vertical Rulers to determine Rater severity/Leniency, and Unexpected Responses to assess rater bias towards the students' writing quality (Sumintono & Widhiarso, 2013; Nur Azizah & Muchlas Suseno, 2023).

RESULTS AND DISCUSSION

Before measuring rater severity, it is crucial to determine the item validity used as parameters for assessing students' writing. To ascertain this, the parameter utilized in the Many-Facet Rasch Measurement is the unidimensionality test (Sumintono & Widhiarso, 2013; Myford & Wolfe, 2004; Huang, 2023). The assumption test criteria used for testing the item's unidimensionality in the MFRM (Many-Facet Rasch Measurement) is the Raw Variance Explained by Measure, with a threshold value greater than 20% (≥ 20%); if it exceeds 40%, it indicates good quality, and if it surpasses 60%, it is considered exceptional (Sumintono & Widhiarso, 2013). Thus, instruments meeting these threshold values are considered to satisfy the unidimensionality requirement or the validity of the construct (Sumintono & Widhiarso, 2013).

Additionally, to identify problematic and incongruent items, the eigenvalue can be examined, which eigenvalue less than 3 indicates that there are no problematic items (Fisher, 2007; Sumintono & Widhiarso, 2013). Figure 1 displays the results of the unidimensionality test for the four items.

		Count	Mean	S.D.
Responses non-extreme estimable	=	240	3,56	0,76
Count of measurable responses	=	240		
Raw-score variance of observations	=	0,579	100.00%	
Variance explained by Rasch measures	=	0,285	49,26%	
Variance of residuals	=	0,294	50,74%	

Figure 1. Item Unidimensionality

The items in assessing students' compositions have a variance explained by measure of 49.26%. This value, being greater than 40%, indicates that all four items exhibit unidimensionality. Additionally, the eigenvalue of 0.285 is significantly below the required value of 3, indicating that there are no problematic items.

The measurement of Person validity is necessary before measuring rater severity. This measurement is needed to determine the accuracy and precision of the raters in assessing students' writing (Misbach & Sumintono, 2014). To assess Person validity in the MFRM, the measurement parameters utilize Outfit mean square (MNSQ), Outfit Z-standard (ZSTD), and Point Measure Correlation with score ranges as shown in Table 3.

Table 3. Person Fit Criteria

Criteria	Score Range
Outfit mean square (MNSQ)	0,5 < *MNSQ* < 1,5
Outfit Z-standard (ZSTD)	-2,0 < *ZSTD* < +2,0
Point Measure Correlation	0,4 < *PT Measure Corr* < 0,85

rater severity 14/07/2023 10:49:00
Table 7.1.1 Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
176	60	2.93	2.93	1.82	.25	1.20	1.0	1.21	1.0	.78	.31	.54	28.3	29.9	2	FTR
219	60	3.65	3.64	-.63	.23	1.13	.8	1.14	.8	.82	.61	.57	37.2	43.3	4	NHR
229	60	3.82	3.81	-1.14	.23	.96	-.1	.94	-.2	1.08	.60	.57	35.6	41.4	1	ARF
231	60	3.85	3.85	-1.25	.23	.58	-2.8	.59	-2.8	1.50	.73	.57	41.1	40.8	3	RHM
213.8	60.0	3.56	3.56	-.30	.23	.97	-.3	.97	-.3		.56					Mean (Count: 4)
22.3	.0	.37	.37	1.25	.01	.24	1.5	.24	1.6		.15					S.D. (Population)
25.7	.0	.43	.42	1.44	.01	.28	1.8	.28	1.8		.18					S.D. (Sample)

Model, Populn: RMSE .23 Adj (True) S.D. 1.23 Separation 5.25 Strata 7.33 Reliability (not inter-rater) .96
Model, Sample: RMSE .23 Adj (True) S.D. 1.42 Separation 6.09 Strata 8.45 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-squared: 104.5 d.f.: 3 significance (probability): .00
Model, Random (normal) chi-squared: 2.9 d.f.: 2 significance (probability): .23
Inter-Rater agreement opportunities: 360 Exact agreements: 128 = 35.6% Expected: 139.9 = 38.9%

Figure 2. Rater Measurement Report

In Figure 2, it can be observed that FTR obtained an outfit MNSQ score of 1.21, an outfit ZSTD score of 1.0, and a point measure correlation of 0.31. NHR achieved an outfit MNSQ score of 1.14, an outfit ZSTD score of 0.8, and a point measure correlation of 0.61. ARF obtained an outfit MNSQ score of 0.94, an outfit ZSTD score of -0.2, and a point measure correlation of 1.08. Lastly, RHM obtained an outfit MNSQ score of 0.59, an outfit ZSTD score of -2.8, and a point measure correlation of 0.73.

Table 4. The Result of Person Fit Order

Number	Rater	Outfit		*PT Measure Corr.*	Person Fit Criteria			Interpretation
		MNSQ	*ZFTD*		*MNSQ*	*ZFTD*	*PT Measure Corr.*	
2	FTR	1,21	1,0	0,31	Fit	Fit	Misfit	Valid
4	NHR	1,14	0,8	0,61	Fit	Fit	Fit	Valid
1	ARF	0,94	-0,2	0,60	Fit	Fit	Fit	Valid
3	RHM	0,59	-2,8	0,73	Fit	Misfit	Fit	Valid

Table 4 presents a summary of the analysis of outfit MNSQ, Outfit ZFTD, and PT Measure Correlation, along with the corresponding person fit criteria and their interpretations. The scores for Outfit MNSQ, Outfit ZFTD, and PT Measure Correlation are compared with the values in Table 3 Person Fit Criteria. The final results of the Person Fit analysis can be found in Table 4, which includes the interpretation of the person fit for each rater.

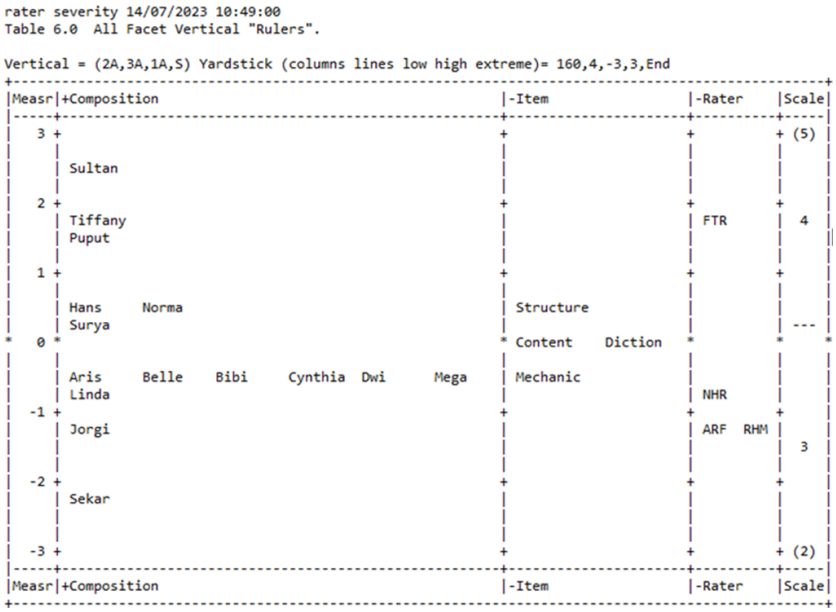

Figure 3. Vertical Ruler

The vertical ruler above represents a form of calibration, where the measured variables are placed on a single scale within the MFRM. The three facets or variables analyzed, namely composition (column 2), item (column 3), and rater (column 3), are placed on the same scale, namely the logit scale (column 1). Meanwhile, column 5 indicates the Likert rating scale established in Table 2. By placing these three facets/variables on the same scale value (scale in logit units), the quality of each facet/variable (composition, item, and rater) can be analyzed or compared based on the logit values.

In the column for rater severity, there is a small gap between rater NHR and raters ARF and RHM, indicating a relatively similar assessment among these three raters. On the other hand, raters ARF and RHM provide almost identical ratings with insignificant differences. There is a considerable gap between rater FTR and the other raters. From these results, it can be concluded that rater FTR has the highest level of severity compared to the other raters (highest logit) (Sumintono & Widhiarso, 2013). While, raters ARF and RHM have the lowest level of severity or leniency compared to others.Furthermore, from the vertical ruler, it can be observed that the distribution of composition is larger than the distribution of items, indicating a mismatch between the difficulty levels of the items and the quality of the compositions. This suggests that the assessment items have a less diverse range of difficulty in measuring the articles. Therefore, it is necessary to add difficult and easy items or, in other words, increase the number of items to more than 4 for the upper and lower parts of the vertical ruler, with difficulty levels adjusted to the logit values. Another possibility is the presence of discrepancies in the raters' assessments, resulting in an uneven distribution.

rater severity 14/07/2023 10:49:00
Table 4.1 Unexpected Responses (14 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	N	Rat	Nu	Composi	N	Item	Sequence
4	4	2.8	1.2	2.4	2	FTR	12	Linda	3	Diction	107
2	2	3.3	-1.3	-2.4	4	NHR	11	Cynthia	2	Structure	222
2	2	3.2	-1.2	-2.3	2	FTR	10	Puput	2	Structure	98
4	4	2.9	1.1	2.3	2	FTR	11	Cynthia	3	Diction	103
4	4	2.8	1.2	2.3	2	FTR	12	Linda	4	Mechanic	108
4	4	2.9	1.1	2.2	2	FTR	11	Cynthia	4	Mechanic	104
2	2	3.1	-1.1	-2.2	4	NHR	14	Jorgi	2	Structure	234
3	3	4.3	-1.3	-2.2	4	NHR	15	Tiffany	1	Content	237
3	3	4.3	-1.3	-2.2	4	NHR	15	Tiffany	3	Diction	239
2	2	3.0	-1.0	-2.1	2	FTR	7	Surya	1	Content	85
2	2	3.0	-1.0	-2.1	2	FTR	9	Norma	3	Diction	95
5	5	3.8	1.2	2.1	4	NHR	2	Hans	3	Diction	187
5	5	3.8	1.2	2.0	4	NHR	2	Hans	1	Content	185
2	2	3.0	-1.0	-2.0	4	NHR	13	Sekar	1	Content	229
Cat	Score	Exp.	Resd	StRes	N	Rat	Nu	Composi	N	Item	Sequence

Figure 4. Unexpected Responses

To identify bias in the raters' assessments, the table that needs to be analyzed from the output of the Many Facet Rasch Measurement is the Unexpected Response table. Figure 4 presents the Unexpected Response or bias in assessments. In the first instance, rater FTR shows bias in the item "diction," giving Linda a score of "4" while the expected score was 2.8. This indicates that Linda received a bonus (residue) of 1.2 points. The second instance shows bias in rater NHR's assessment of Cynthia's composition on the item "structure." The score given is 2, whereas the expected score was 3.3, resulting in a reduction of -1.3 points for Cynthia. These results reveal that both Linda and Cynthia received biased assessments.

In cumulative terms, bias or unexpected responses occurred 7 times for rater FTR and 7 times for rater NHR. The unexpected responses occurred 5 times in the "diction" item, 3 times in "structure," 2 times in "mechanic," and 4 times in "content." It can be concluded that both rater NHR and FTR require special attention regarding the accuracy of their assessments. Thus, it may be necessary to conduct retraining to align the perceptions among raters. Figure 4 also suggests that rater ARF and RHM have relatively accurate assessment accuracy. In other words, these raters have similar and consistent perceptions in applying the assessment rubric according to the criteria specified in the rubric.

Despite significant severity differences among instructors on the vertical ruler, outfit, and infit values for raters scoring between 0.59 to 1.21 indicate that internal consistency was observed in each instructor's ratings. This

finding is promising, as previous researchers, such as McNamara (1996), considered random errors in raters with respect to internal consistency to be more detrimental than systematic and explainable rater effects. Therefore, what is needed is to ensure consensus in perceptions and retraining for raters with excessively high severity (Fahim & Bijani, 2011).

CONCLUSION

The findings indicate that there are variations in rating behavior among raters, leading to dissatisfaction among students graded by more severe instructors. As a result, the study aims to promote standardization in the composition ratings. A common approach to achieving this is through training sessions, where instructors adhere closely to a predefined set of rubric criteria and assess essays accordingly. The outcomes reveal whether instructors' interpretations align with those of other raters, thus ensuring consistency in rating criteria interpretation. The main objective of such training is to minimize variability and randomness in overall severity or leniency.

Meanwhile, the occurrence of score bias further confirms that rater severity is directly proportional to score bias. The same rater (the most severe one) tends to provide less accurate scores. Score bias occurs when severe raters pay less attention to the established criteria (traits) for each item, making it challenging for them to give objective scores. Other factors may also play a role in score bias, such as subjectivity to criteria or personal preferences. Therefore, it is recommended to prevent subjectivity to criteria by conducting retraining to avoid rater severity. To avoid the influence of personal preferences, objective evaluators who do not have a direct relationship with the students are needed.

Furthermore, the study also found that having more than four items/criteria (content, diction, structure, and mechanic) is essential to achieve a more diverse distribution of item difficulty and effectively measure students' writing abilities. A limited number of items may result in an uneven distribution, focusing on a narrower range. Therefore, it is necessary to include additional difficult and easy items with adjusted difficulty levels to ensure a balanced spread of item difficulty. This can be achieved by breaking down the four existing items into more specific ones or adding new items such as cohesion and coherence.

REFERENCE

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. SAGE Open, 9(1). https://doi.org/10.1177/2158244018822377

Bond, T., & Fox, C. (2015). Applying the Rasch Model. Routledge. https://doi.org/https://doi.org/10.4324/9781315814698

Bond, T.G., & Fox, C. M. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd ed.). Awrence Erlbaum Associate.

Cahyono, B. Y. (2019). Teaching English as a foreign language in Indonesia: Past, present, and future directions. TEFLIN Journal, 30(2), 141-156.

Erguvan, I. D., & Aksu Dunya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. Language Testing in Asia, 10(1). https://doi.org/10.1186/s40468-020-0098-3

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. Iranian Journal of Language Testing, 1(1), 1–16.

Fisher, W. (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions, 1.

Huang, H. Y. (2023). Modeling rating order effects under item response theory models for rater-mediated assessments. Applied Psychological Measurement, 47(4), 312–327. https://doi.org/10.1177/01466216231174566

Jacobs., Holly. L., Stephen, A., Zingkgraf., Deanne. R., Wormuth, V., Faye, H., Jane, B., H. (1981). Testing ESL Composition: A Practical Approach. Newbury House Publishers, Inc.

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. Reading and Writing, 35(10), 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Lim, G. S. (2009). Prompt and Rater Effects in Second Language.

Linacre, J. M. (2002). KR-20/Cronbach alpha or Rasch person reliability: Which tells us the truth? Rasch Measurement Transactions, 11, 580–581.

Linacre, J. M. (2006). Many-facet Rasch measurement. In Facets Rasch Measurement.

Linacre, J. M. (2020). What do infit and outfit mean-square and standardized mean? Rasch Measurement Transaction, 16, 878.

Maier, K. S. (2001). A Rasch hierarchical measurement model. Journal of Educational and Behavioral Statistics, 26, 307–331. https://doi.org/https://doi.org/10.3102%2F10769986026003307

Misbach, I. H., & Sumintono, B. (2014). Pengembangan dan Validasi Instrumen "Persepsi Siswa Tehadap Karakter Moral Guru" di Indonesia dengan Model Rasch. PROCEEDING Seminar Nasional Psikometri, 148–162.

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.

Noor, N., Beram, S., Huat, F. K. C., Gengatharan, K., & Mohamad Rasidi, M. S. (2023). Bias, Halo effect and Horn effect: A systematic literature review. International Journal of Academic Research in Business and Social Sciences, 13(3). https://doi.org/10.6007/ijarbss/v13-i3/16733

Nur Azizah, Muchlas Suseno, B. H. (2023). Penilaian Menulis Menggunakan Many Fcets Rasch Measurement (MFRM) Pengaplikasian Software FACETS Dalam Validasi

DOI: http://dx.doi.org/10.30998/scope.v8i1.19432

Instrumen Penilaian Menulis Serta Analisis Penilai dan Karya Tulis.pdf (p. 93). Program Pasca Sarjana Universitas Negeri Jakarta.

Rahayu, S., & Widiati, U. (2019). Challenges in writing academic texts in English: A case study in an Indonesian university. Journal of Applied Linguistics and Language Research, 6(4), 144-158.

Setiawan, H., & Hartoyo, N. S. (2020). The challenges of translating idiomatic expressions: A case study of Indonesian EFL students. 11(1), 153-166.

Sumintono, B. & Widhiarso, W. (2013). Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial. Trim Komunikata Publishing House.

Tan, S. (2013). V. of an A. R. S. for W. A. R. M. A. (2013). Validation of an Analytic Rating Scale for Writing: A Rasch Modeling Approach. Tabaran Institute of Higher Education. Iranian Journal of Language Testing, 3(1).

Tanaka, M. S. J. R. (2023). Impact of self-construal on rater severity in peer assessments of oral presentations. Assessment in Education: Principles, Policy & Practice, 30(2), 203–220. https://doi.org/doi.org/10.1080/0969594X.2023.2189564

Uto, M. (2022). A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. Behavior Research Methods, October. https://doi.org/10.3758/s13428-022-01997-z

Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. Appl Psychol Meas, 43(2), 159–171. https://doi.org/10.1177/0146621618789391

APPENDIX

Syntax Coding

TITLE = "rater severity"

Facets = 3

Inter-rater = 1

Positive = 2

Non-centered = 1

Pt-biserial = Measure

Yard = 160,4

Model = ?B,?B,?,R5

Unexpected = 2

Vertical = 2L,3A,1L

Arrange = m,F,N

Zscore = 1,2

Labels =

1, Rater

1 = ARF

2 = FTR

3 = RHM

4 = NHR

*

2, Composition

1 = Belle

2 = Hans

3 = Aris

4 = Dwi

5 = Sultan

6 = Bibi

7 = Surya

8 = Mega

9 = Norma

10 = Puput

11 = Cynthia

12 = Linda

13 = Sekar

14 = Jorgi

15 = Tiffany

*

3, Item

1 = Content

2 = Structure

3 = Diction

4 = Mechanic

*

Data=

1,1,1-4,4,3,3,4

1,2,1-4,3,3,3,3

1,3,1-4,3,3,3,3

1,4,1-4,3,3,3,3

1,5,1-4,5,4,5,5

1,6,1-4,3,3,3,4

1,7,1-4,5,4,5,5

1,8,1-4,4,4,4,4

1,9,1-4,4,4,4,4

1,10,1-4,5,4,4,4

1,11,1-4,4,4,4,4

1,12,1-4,4,4,4,4

1,13,1-4,3,3,3,3

1,14,1-4,4,4,4,4

1,15,1-4,5,4,5,5

2,1,1-4,3,2,3,3

2,2,1-4,3,3,3,3

2,3,1-4,2,3,3,3

2,4,1-4,3,2,3,3

2,5,1-4,3,4,3,4

2,6,1-4,2,2,3,3
2,7,1-4,2,2,3,3
2,8,1-4,3,2,3,3
2,9,1-4,3,3,2,3
2,10,1-4,3,2,3,3
2,11,1-4,3,3,4,4
2,12,1-4,2,3,4,4
2,13,1-4,3,2,3,3
2,14,1-4,3,3,3,3
2,15,1-4,4,4,3,3
3,1,1-4,4,4,4,4
3,2,1-4,4,4,4,4
3,3,1-4,4,4,4,4
3,4,1-4,4,4,4,4
3,5,1-4,5,5,5,4
3,6,1-4,4,4,4,4
3,7,1-4,4,3,3,3
3,8,1-4,4,4,4,4
3,9,1-4,4,4,4,4
3,10,1-4,5,5,4,4
3,11,1-4,4,3,3,4
3,12,1-4,3,3,3,3
3,13,1-4,3,3,3,3
3,14,1-4,3,3,3,3
3,15,1-4,5,4,5,5
4,1,1-4,3,3,3,4
4,2,1-4,5,4,5,5
4,3,1-4,4,4,4,4
4,4,1-4,4,4,4,4
4,5,1-4,5,4,4,4
4,6,1-4,4,4,4,4
4,7,1-4,4,4,4,4
4,8,1-4,3,3,3,3
4,9,1-4,4,4,4,4
4,10,1-4,5,4,5,5
4,11,1-4,3,2,3,3
4,12,1-4,3,3,3,3
4,13,1-4,2,3,3,3
4,14,1-4,3,2,3,3
4,15,1-4,3,4,3,4

 DOI: http://dx.doi.org/10.30998/scope.v8i1.19432