Identifying Rating Interaction Effects among Indonesian University Lecturers in Assessing English Essays: A Many Facet Rasch Measurement Analysis

Muhammad Affan Ramadhana [1, 3,*] Bambang Sumintono [1] and Zulfa Sakhiyya [2]

[1] Universitas Islam Internasional Indonesia, Depok, West Java, Indonesia
[2] Universitas Negeri Semarang, Semarang, Central Java, Indonesia
[3] Politeknik Dewantara, Palopo, South Sulawesi, Indonesia
*affanramadhana@live.com

Abstract. Despite the use of standard analytical rubrics, assessing English as a Foreign Language (EFL) writing often remains susceptible to inconsistency and rating interaction effects. To further explore that issue in Indonesian higher education context, this study employs a Many-Facet Rasch Measurement (MFRM) analysis to identify interaction effect from raters' backgrounds toward rubric criteria in English essay assessment. This study included 36 Indonesian university lecturers who assessed English essays using the *ESL Composition Profile* analytical rubric. The analysis specifically explored interaction effects between rater background and rubric criteria. The results show different assessment behaviors of lecturers based on their academic qualifications, field of study, professional rank, and gender. This study is expected to offer practical insights for assessment committees to minimize potential biases, especially when designing multi-rater English writing assessment programs.

Keywords: writing assessment, interaction effect, rating behavior, rater background, distal facets.

1 Introduction

One important issue in writing assessment is that it relies on human judgment. Because writing assessment is a form of rater mediated assessment, the rater becomes a central component of the measurement process. Therefore, one key concern is whether the rater is free from biases or interaction effects. Such effect may occur at the individual level, but it may also arise from broader background factors attached to the rater, such as educational background, professional experience, or other personal attributes.

Identifying rater interaction effect is critical because, in writing assessment, particularly in the context of English as a Foreign Language, it poses a serious threat to validity. Even when a well-designed rubric is used, biased rater judgment can distort

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_18

assessment results. In such cases, the scores produced may no longer accurately represent students' writing ability, and the validity of the assessment is compromised.

Previous empirical studies have shown that rater bias, often referred to as rater effects, tends to be systematic rather than random [1]. This means that it does not occur by chance, but is influenced by identifiable factors related to the rater's background or characteristics. These systematic sources of effects may be caused by differences in training, experience, disciplinary orientation, or long-established assessment habits. Therefore, examining rater interaction effect is essential for understanding and improving the quality and fairness of writing assessment.

To address this issue, a specific analytical approach is required to detect rater interaction effect that occurs when raters assess English as a Foreign Language (EFL) writing. In the Indonesian context, studies that focus on detecting rater interaction effect in writing assessment are still limited. Therefore, to examine and identify that effect, particularly systematic interaction effect related to rater background, this study applies Many Facet Rasch Measurement (MFRM) analysis.

In the context of writing assessment, MFRM allows for the detection of systematic rater effects by examining interactions among multiple facets. These facets include individual raters, groups of raters based on background characteristics, the essays being rated, and the rubric criteria used for scoring. Through this interaction analysis, it becomes possible to observe patterns of rater severity or leniency at both the individual and group levels, and to examine how raters with different backgrounds apply specific rubric criteria.

Previous studies, including those conducted in EFL writing assessment contexts, have demonstrated that MFRM is effective in identifying rater bias/interaction effect [1, 2]. Beyond detection, MFRM also allows for score adjustment, producing estimates of writing quality that account for rater severity and the difficulty of rubric criteria [3, 4]. As a result, the final scores generated through MFRM can be interpreted as fairer and more accurate representations of students' writing ability.

Despite its established use in language assessment more broadly, the application of MFRM in the Indonesian context, particularly for EFL writing assessment at the higher education level, remains relatively limited. This is the case even though MFRM has become a standard analytical approach in many areas of language assessment. One possible reason for this limited use is the practical complexity of implementing MFRM analysis, which requires careful design, sufficient data, and technical expertise. Moreover, conducting this type of analysis is demanding in local and context sensitive settings because it requires active involvement from lecturers. Raters must be brought together within a structured research activity to conduct scoring tasks, and in many cases, they also need to participate in rater training. These requirements make such studies difficult to implement in higher education contexts in Indonesia. As a result,

research that closely examines rater behavior and interaction effect in writing assessment remains limited.

For this reason, the present study aims to explore and identify the types of rater interaction effect that occur among lecturers in Indonesian higher education when they assess English essays. In this study, English is treated as a foreign language, which reflects the local instructional context. The focus is on identifying systematic rater interaction effects that may occur when lecturers apply scoring rubrics in assessing students' writing.

More specifically, this study formulates five research questions to investigate rater interaction effect across different groups of lecturers. The analysis focuses on interaction effect between lecturer groups based on background characteristics when they assess English essays using a writing rubric.

1. Is there a significant interaction effect between rater gender and the rubric criteria used in rating?
2. Is there a significant interaction effect between raters' highest educational qualification and the rubric criteria used in rating?
3. Is there a significant interaction effect between raters' academic field and the rubric criteria used in rating?
4. Is there a significant interaction effect between raters' professional rank and the rubric criteria used in rating?
5. Is there a significant interaction effect between raters' teaching homebase and the rubric criteria used in rating?

2 Conceptual Framework

The framework for this study is based on the Eckes' model [5]. In rater mediated assessment, there are two broad types of factors that influence the scores assigned by raters. The first type is referred to as *proximal facets*. These are factors that directly determine how scores are assigned. Proximal facets include the rater, the examinee or text being assessed, the scoring criteria or rubric, and the task or performance produced by the examinee. These elements form the core components of the assessment process, as they are directly involved in the act of scoring.

In addition to proximal facets, the model also recognizes the role of *distal facets*. Distal facets refer to background conditions that influence assessment indirectly. These include characteristics and backgrounds of the rater, characteristics of the examinee, and contextual factors such as assessment setting, timing, and environmental conditions. Although these factors do not directly constitute the scoring, they can shape how raters interpret performances and apply scoring criteria.

In this study, the focus is placed on distal facets related to the rater. Specifically, the study examines background conditions that may influence rater behavior. Five types of rater background variables were collected and used to group raters accordingly. These background-based groupings form the central focus of the analysis in this study.

This focus is important because variation among raters should not be treated merely as random noise. Rater variation is structured and systematic rather than purely random [6]. By applying the MFRM model, interactions among multiple facets can be explicitly modeled. Distal facets can be specified as separate facets within the model, allowing their interactions with other facets to be estimated. In this study, these interactions are examined in relation to rubric criteria, enabling a detailed analysis of how rater background conditions influence scoring behavior across different aspects of writing.

3 Method

3.1 Participants/Raters

The participants for this research are university lecturers in several provinces of Indonesia, who have experiences in teaching and assessing English essays written by university students. The criteria for participant inclusion are: 1) registered as permanent lecturers in accredited higher education institutions in Indonesia; and 2) experienced and able to assess English essays written by university undergraduate students. Consequently, all lecturers registered in English-language related programs are eligible to participate. However, lecturers who have educational background related to English-language studies but currently teaching English (for Specific Purposes) in other programs were also invited to participate.

Participants in this study are 36 Indonesian university lecturers specializing in English, all of whom held relevant postgraduate degrees in English language education, linguistics, or literature studies.

Prior to the assessment, all participants were invited to attend rater training, which focused on understanding descriptors and scoring using the rubric. Rater trainings were conducted in three formats: virtual sessions, onsite sessions, and asynchronous self-training. Among the participants, 12 lecturers chose to attend the virtual training, 11 participated in the onsite sessions, and 13 opted for self-training where they independently reviewed and learned the provided rater training materials.

3.2 Essays

In this study, the essays to be rated for assessment is collected from the International Corpus Network of Asian Learners of English (ICNALE) developed by Dr. Shin Ishikawa of Kobe University, Japan [7, 8]. ICNALE is a comprehensive international learner corpus that contains more than 10,000 essays and speeches from college students across Asian countries and regions, as well as data from native English

speakers [7]. More specifically, ICNALE uniquely focuses on learners from Asia from both English as a Second Language (ESL) regions (e.g., Hong Kong, Pakistan, the Philippines, Singapore) and English as a Foreign Language (EFL) regions (e.g., China, Indonesia, Japan, Korea, Taiwan, and Thailand), which provides corpus study of both learners' English and the diverse varieties of English across Asia.

For this study, 45 essays were collected from the Written Essay corpus (Ishikawa, 2023, 2024b) which contains 200-300 words essays. This study focuses only on the topic "It is important for college students to have a part-time job" that is written by college students from EFL regions. The students' proficiency levels range from B1_1 (low), B1_2 (mid), and B2 (high).

3.3 Rubric

The raters were asked to assign rating to each essay for each criterion based on Jacobs' et al. [9, 10] *ESL Composition Profile*. The rubric consists of five criteria: *Content*, *Organization*, *Vocabulary*, *Language Use*, and *Mechanics*. The reason for choosing this rubric is based on its stability and consistent use in many second language (L2) writing research e.g. [11, 12] and also used in rated version of the essays in the ICNALE [13].

However, a Rasch measurement analysis with multiple facets revealed that the 20-point rating scales within the original *ESL Composition Profile* rubric demonstrated poor performance [9]. Therefore, this research will use a revised version of the rubric which collapses the 20-point scale into 4-point scale. Therefore, each rubric criterion is defined by four level descriptors, represented on a 4-point scale. The revised version is proved to improve to rater consistency and increased raw agreement among raters, particularly in terms of performance-level descriptors [9].

3.4 Rating Assignment Design

In distributing the essays to be rated, the rating design employed a partially crossed design [5], also referred to as sparse rating assignment design [14]. In this design, only several essays were given to each rater, yet there will be overlapping assignments to ensure connectedness among raters in the whole design. After 36 lecturers completed the ratings, the rating design was all connected partially without any disconnected subsets. Figure 1 shows visualization of all rater connections, where the light blue nodes represent raters and red nodes represent essays.

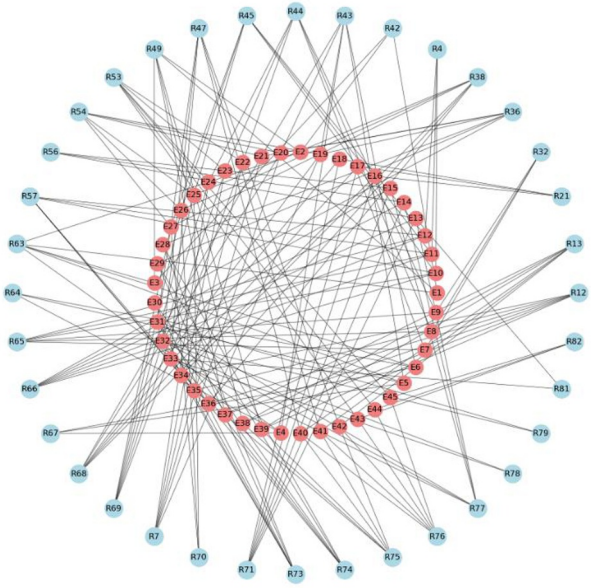


Fig. 1. Rater Connection Network Visualization

3.5 MFRM Model

The MFRM analysis was conducted using Minifac64 software [15]. Considering the nature of multiple facets of data, such as raters, essays, and rubric criteria, Rasch measurement theory is particularly useful with this complexity. Moreover, given the partially crossed design of rating assignment in this research, there are a lot of missing data to be dealt with. By using MFRM analysis, sparse data from all the potential ratings can still be estimated, because MFRM demonstrates strong robustness to missing data [16, 17].

To do the analysis, the model specification includes three main facets: raters, essays, and rubric criteria. Additionally, five dummy facets were also added for bias/interaction analysis between raters' background and rubric criteria. The dummy facets here are the raters' demographic properties such as gender, educational qualifications, educational field, professional rank, and teaching homebase.

The exact equation of MFRM with rater background subgroup interaction by rubric criterion model is as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \gamma_g - \varphi_{gi} - \tau_k \quad (1)$$

where

p_{nijk} = probability of essay n receiving a rating of k from rater j on criterion i,

p_{nijk-1} = probability of essay *n* receiving a rating of *k* – 1 from rater *j* on criterion *i*,
θ_n = quality of essay *n*,
β_i = difficulty of criterion *i*,
α_j = severity of rater *j*,
γ_g = effect of rater background (dummy facets) subgroup *g*,
φ_{gi} = effect of the interaction between rater background subgroup *g* and criterion *i*,
τ_k = difficulty of receiving a rating of *k* relative to *k* – 1.

4 Results

From the MFRM model estimation, the mean score (measure) of the rater facet is 0.00 logit (SD = 1.21), mean score (measure) of the criteria facet is also 0 logit (SD = 0.17), while mean score (measure) of the essay facet is 1.54 logit (SD = 1.40). For the rater facet, the model produced a population RMSE of .43 and a standard deviation of 1.13. The rater facet obtained a separation index of 2.62 and a strata value of 3.83. The reliability value of 0.87 indicated that the model could distinguish raters according to their location along the severity continuum. The fixed-effects chi-square is 413.3 with 35 degrees of freedom and significance level of 0.00, which indicates that the raters have significantly different behavior in rating. In total, among 1,390 rating decisions recorded, there are 593 exact agreements (42.7%), which closely approximated the expected agreement of 575.3 (41.4%). The following Wright Map (Fig. 2) shows how the three facets are located in the measure continuum.

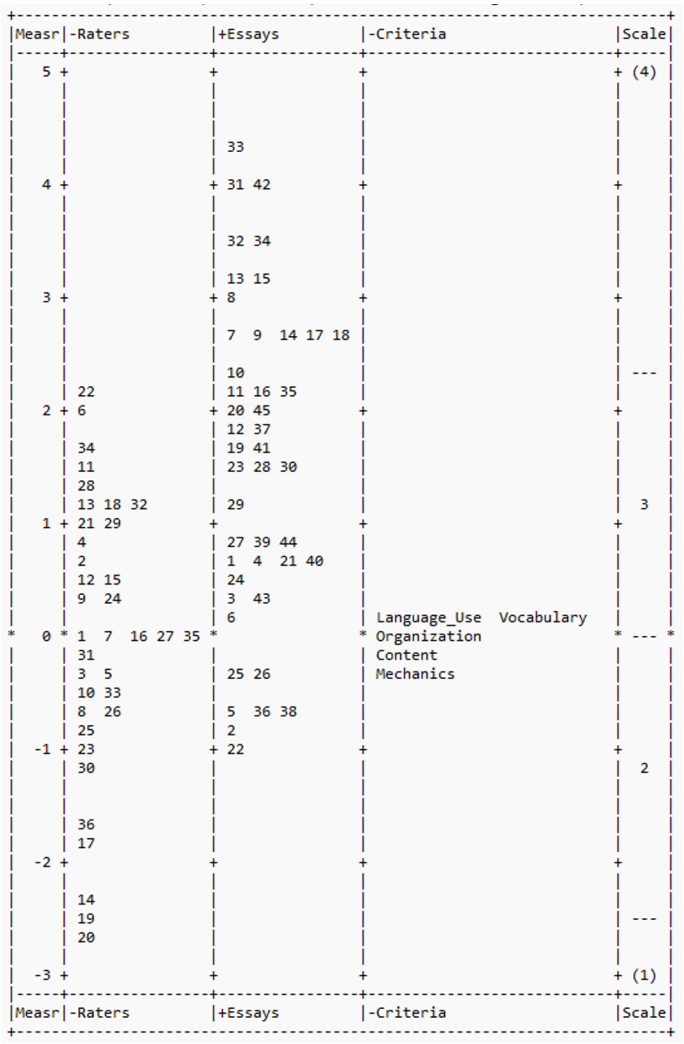


Fig. 2. Wright Map of Three Main Facets

Then, the analysis of rater background interaction with rubric criteria was run to provide a detailed view of how raters' personal characteristics shaped their interpretation of the rubric. To conduct this analysis, dummy facets which are the raters' demographic properties such as gender, educational qualifications, educational field, professional rank, and teaching homebase, were run one by one for their bias/interaction with rubric criteria facet. These dummy facets, representing proximal facets [5], allowed the analysis to determine whether different rater groups as defined by their background properties showed systematic differences in how they assigned scores across the five rubric criteria.

The analysis of interaction effects between rater background towards rubric criteria is based on the bias size and the Rasch-Welch t-value. Interactions effect in this context typically arises when a rater shows a response pattern that deviates from what is expected by the Rasch model. The Welch t-test value is used to determine whether this deviation is significant enough to be considered systematic interactions effect. Generally, if the Rasch-Welch t-value value above +2 or below –2 and indicates statistical significance (p-value < 0.05), it suggests the presence of interactions effect in the data [18].

4.1 Interaction Effect between Raters' Gender and Rubric Criteria

First interaction effect analysis on raters' background factors is between raters' gender towards rubric criteria. Based on their gender, raters are divided into two groups. First group is male raters, with 14 members, and second group is female, with 22 members. The following line chart shows how raters from each of the two groups behave in rating the five rubric criteria.

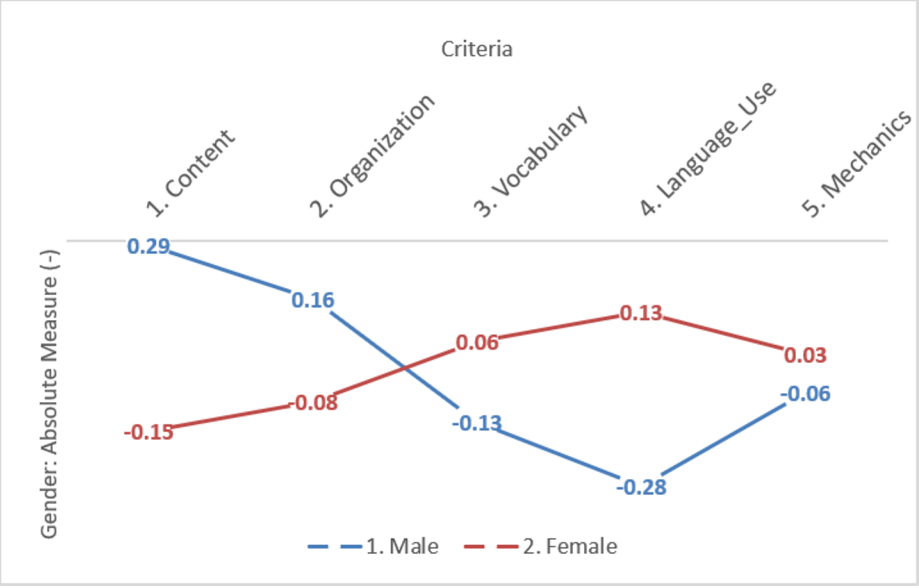


Fig. 3. Rating Interaction Effect on Gender towards Rubric Criteria

As shown in Figure 3, bias/interaction analysis between raters' gender and rubric criteria produced interesting contrasts. Male raters show severity in *Content* and *Organization* but show leniency in *Vocabulary*, *Language Use*, and *Mechanics*. Female raters displayed the opposite pattern, with leniency in *Content* and *Organization* and severity in the other three criteria. Their graphs appear almost as inverse shapes. However, as shown in the vertical scale in the chart, differences in rating behavior here are less than one logit with no significant biases occurred. Here, the most extreme difference occurred only within male raters themselves when they assigned severe

scores in *Content* compared to lenient scores in *Language Use*, with 0.57 logits in target contrast and Rasch-Welch *t* test 1.86 (Prob. 0.0650). Thus, it can be concluded that although there are notable differences in rating severity between males and females, the differences are so small and probably happen only by chance.

4.2 Interaction Effect between Raters' Qualifications and Rubric Criteria

Second interaction effect analysis was conducted to explore the interaction between raters' educational qualifications towards rubric criteria. Indonesian higher educational standard regulate that university lecturers must have a minimum qualification equivalent to master's degree. Here, raters are grouped based on their highest educational degree, doctoral degree holders and master's degree holders. The following line chart shows how raters from each of the two groups behave in rating the five rubric criteria.

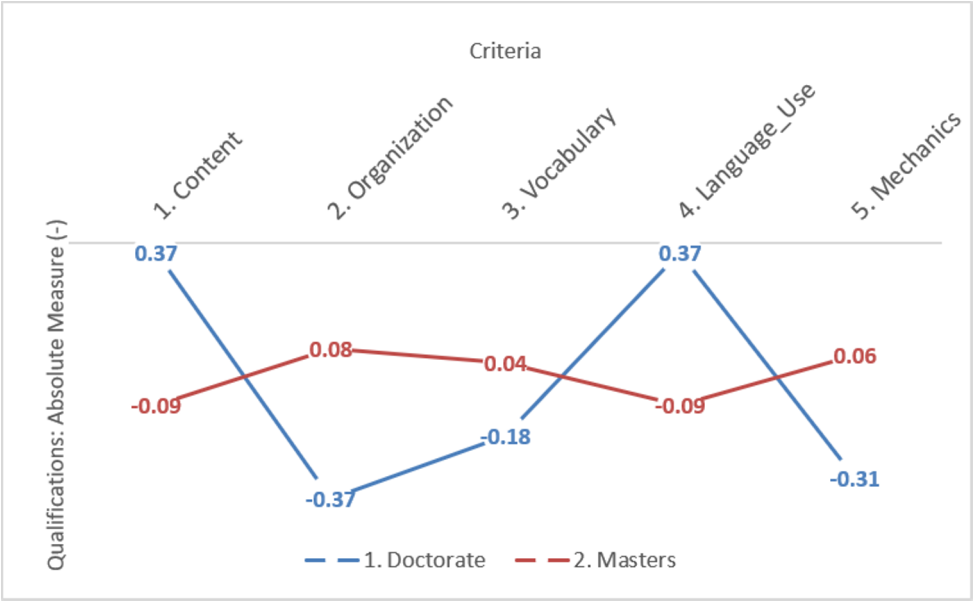


Fig. 4. Rating Interaction Effect on Qualifications towards Rubric Criteria

Across all raters' educational qualifications variables, notable contrasts emerged. Raters with doctoral degrees were very severe in *Content* and *Language Use*, but unexpectedly lenient in *Organization*, *Vocabulary*, and *Mechanics*. Their line graph shows sharp rises and drops across the five criteria, producing a highly fluctuating pattern. On the contrary, raters with master's degrees show a more stable line, with smaller fluctuations across criteria.

Figure 4 shows that the most extreme differences when raters with doctoral degrees assigned severe scores in *Content* compared to lenient scores in *Organization*, with 0.74

logits in target contrast and Rasch-Welch *t* test 1.82 (Prob. 0.0724), which is not significant. Another extreme difference also occurred when they assigned severe scores again in *Language Use* after being lenient in *Organization,* with -0.74 logits in target contrast and Rasch-Welch *t* test -1.86 (Prob. 0.0672), which is also not significant. Therefore, although the contrast pattern between raters with doctoral degrees and raters with master's degrees, this finding shows that no significant interaction effects occurred in this analysis.

4.3 Interaction Effect between Raters' Educational Field and Rubric Criteria

Next interaction effect analysis is based on raters' educational field. In Indonesian higher educational system, professional lecturers who teach English language mostly graduated from either of two cluster of educational fields: English language education; or English literature and linguistics. In this analysis, the raters are grouped based on their field of study at their highest educational level. To simplify the analysis, it did not account for raters who shifted their field from their previous education in undergraduate and/or master's degree level.

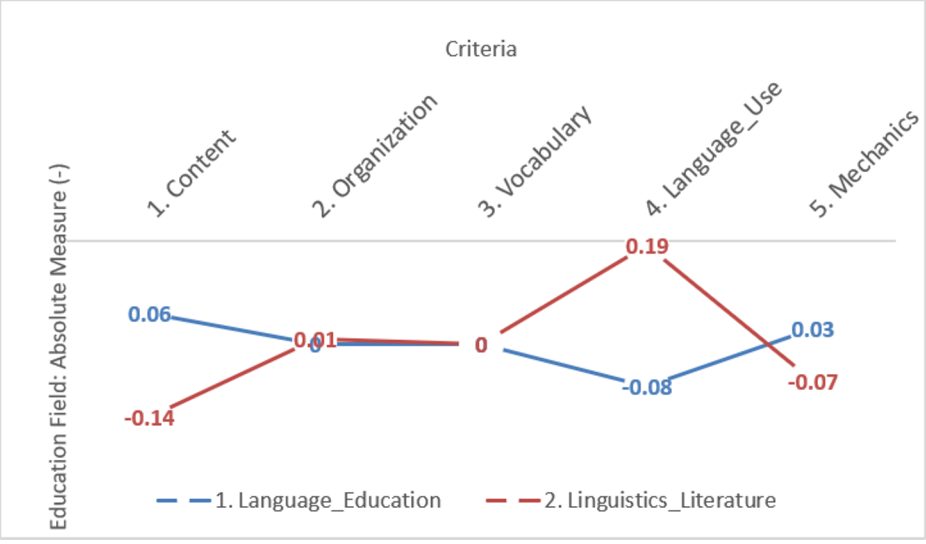


Fig. 5. Rating Bias on Education Field towards Rubric Criteria

The analysis shows that field of study might also influence scoring patterns. Raters from language education backgrounds were seen more severe in *Content* but noticeably lenient in *Language Use*. In contrast, those from linguistics and literature backgrounds were lenient in *Content* but severe in *Language Use*. Their line graphs show almost mirror-image patterns when comparing these two groups. These patterns suggest that raters draw on disciplinary habits when assessing the essays. Raters from language education backgrounds prioritize meaning and message, while raters from linguistics and literature backgrounds prioritize form.

However, by looking at the numbers, the difference between the two behaviors when rating *Content* is very small and not significant. Rasch-Welch *t* test 0.70 (Prob. 0.4846) and target-contrast 0.19 logit shows that there is almost no difference in meaningful scale when *Content* of the essays was scored by either raters from language education backgrounds or raters from linguistics and literature backgrounds. Similar numbers appear when those raters rate the *Language Use* aspect, where the Rasch-Welch *t* test -1.02 (Prob. 0.3117) and target-contrast -0.27 logit. Therefore, while the contrast pattern between raters with different fields of study may indicate different behavior and sensitivity toward some assessment criteria, this finding shows that it is not significant and not enough to support that argument.

4.4 Interaction Effect between Raters' Professional Rank and Rubric Criteria

Another interaction effect analyzed is between raters' professional rank and the rubric criteria. Professional rank was treated as a dummy facet consisting of four hierarchical levels: Instructor (entry-level teaching staff), Lecturer, Assistant Professor, and Associate Professor.

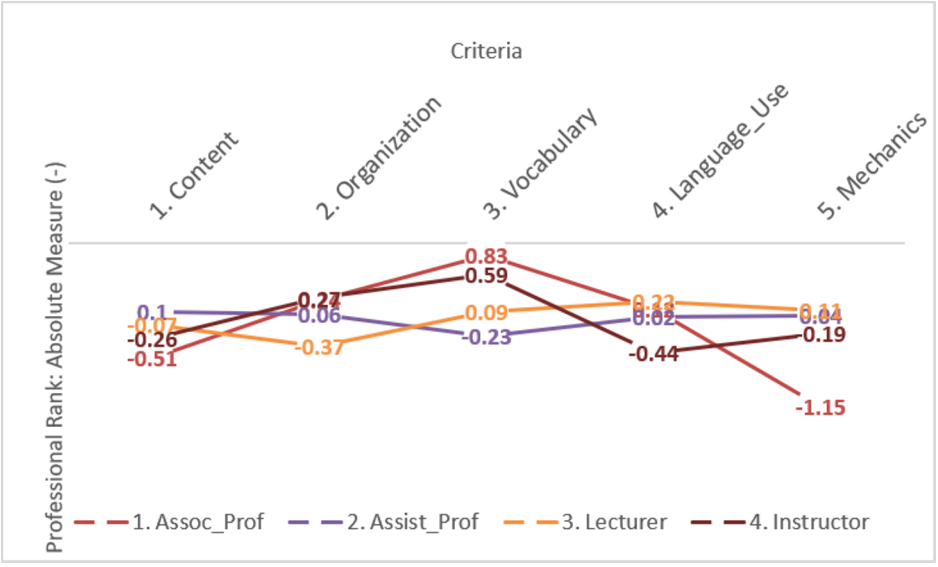


Fig. 6. Rating Bias on Professional Rank towards Rubric Criteria

The first notable pattern appears at the level of Associate Professors, whose rating behavior shows visible fluctuations across criteria. When scoring *Content*, the Associate Professor group shows target-measure of -0.51 logit, which indicates slightly lenient behavior towards this criterion. However, when evaluating *Vocabulary*, the severity increases, with a target-measure of 0.83 logit. This pattern then reverses when

scoring *Mechanics*, where the target measure drops sharply to -1.15 logit. This shift indicates that Associate Professors have different severity levels across criteria.

A similar fluctuation occurs among Instructors, the entry-level rank of Indonesian university lecturer system. When Instructors assess *Content*, they appear lenient, with a target measure of -0.26 logits. In contrast, when scoring *Vocabulary*, their severity increases, with a target measure of 0.59 logits. The unexpected behavior appears again when they assess *Language Use*. Instead of maintaining or increasing severity, Instructors shift back toward leniency, with a target measure of -0.44 logits. This contrast behaviors from Instructors results in a target contrast of 1.04 logits between *Vocabulary* and *Language Use*. The Rasch–Welch t statistic for this comparison is 2.33 (Prob. 0.0242), which indicates that this shift is statistically significant. The Instructors thus demonstrate a significant interaction effect, where they are more severe on *Vocabulary* yet significantly more lenient on *Language Use*.

A second significant interaction effect case is when *Vocabulary* is rated by Assistant Professors compared with Instructors. *Vocabulary* was rated leniently with a target measure of -0.10 logits by Assistant Professors. However, when rated by Instructors, the target measure increases in severity to 0.72 logits. In this case, the target contrast is -1.82 logits and the Rasch-Welch t statistic is -2.36 (Prob. 0.0241), confirming the significance of this difference. This indicates that Instructors are significantly more severe than Assistant Professors when scoring *Vocabulary*.

For the other rank levels, such as Lecturers and Assistant Professors, most comparisons did not yield significant interaction effects. While minor fluctuations appear in the line chart, these do not surpass the statistical thresholds required to indicate interaction effect within the Rasch model. Overall, the analysis shows that the only statistically meaningful bias interactions appear among Instructors (entry-level rank). This finding suggests that Instructors, as early-career lecturers with limited experience, may display less stable scoring patterns when faced with complex writing assessment tasks.

4.5 Interaction Effect between Raters' Teaching Homebase and Rubric Criteria

The next dummy facet examined in the interaction effect analysis is the teaching homebase of the raters, that is, the academic department or study program in which each lecturer is professionally homebased. In the Indonesian higher-education context, English course is delivered across a wide range of departments, and lecturers responsible for teaching English do not always belong to English-specific department or units. For this reason, the homebase information of the raters are divided into three categorical groups.

The first group consist of lecturers affiliated with English Language Education programs, the largest and most common category nationwide. These programs focus on

preparing graduates in English language teaching, and the raters in this group hold academic positions in departments whose core mission centers on English teaching and pedagogy. A total of 23 raters belonged to this group.

The second group includes lecturers based in English Literature and/or English Language programs, sometimes labeled "Bahasa dan Sastra Inggris" in Indonesian universities. Other programs might include vocational education program, named as English for Business and Professional Communication. These programs emphasize English as an academic discipline through applied linguistics, language use, and literary studies rather than teacher education. Six raters in the dataset belonged to this category.

The third group consists of lecturers who teach English but are based in non-English departments. This group includes lecturers assigned to departments such as Information Systems, Accounting, Sharia Banking, Informatics Engineering, Elementary Teacher Education, and Automotive Engineering, among others. While these lecturers regularly teach English courses, their academic homebase lies outside English-focused programs. This group consist of seven raters, and for this analysis here labeled as 'Others.'

Using these three groups, the interaction effect between teaching homebase and rubric criteria was analyzed. The line chart displays the absolute measures for each group across the five rubric criteria.

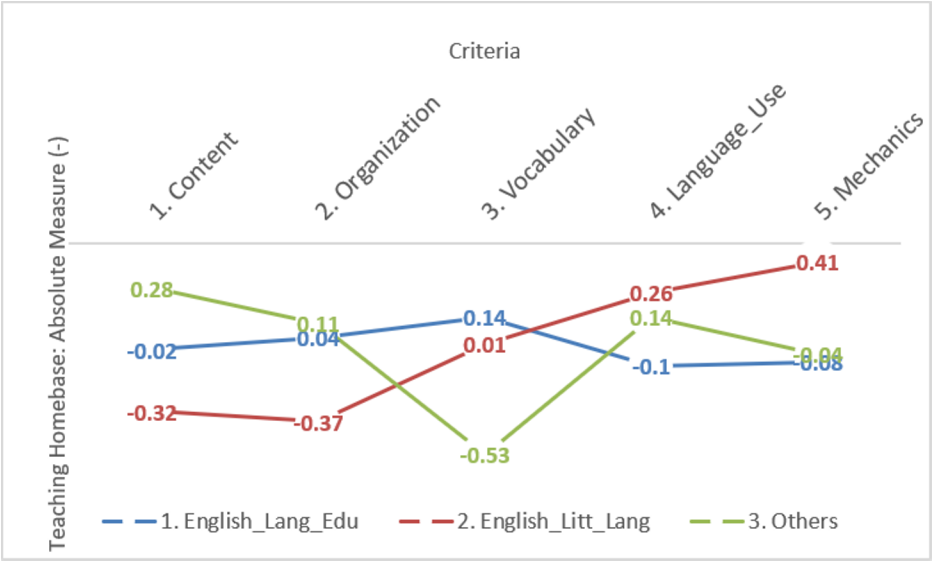


Fig. 7. Rating Bias on Teaching Homebase towards Rubric Criteria

Raters in the English Literature and Language group display a fluctuating but upward-moving pattern. Their severity levels are lenient on *Content*, then a little more lenient when assessing *Organization*, and then increasing severity gradually across

Vocabulary, *Language Use,* and *Mechanics*. This pattern indicates that they become increasingly severe when evaluating aspects more closely tied to linguistic accuracy and disciplinary expectations.

The most distinctive and fluctuating pattern appears in the non-English departments (Others) group. These raters are the most severe in evaluating *Content*, which aligns with their disciplinary orientation. This behavior reflects a high valuation of idea development and subject-matter clarity. These raters are also severe in assessing *Organization*, but when assessing *Vocabulary*, the severity drops sharply and. It means that these raters are highly lenient on *Vocabulary*. Then they are severe again on *Language Use*, before becoming slightly lenient again on *Mechanics*. This inconsistent pattern indicates that this group prioritizes *Content* more strongly than linguistic precision. This behavior is consistent with many non-English academic contexts where English functions primarily as a medium for conveying disciplinary knowledge.

In contrast to the other two groups, raters based in English Language Education show a stable, plateau-like pattern. Their severity remains close to the logit zero line across all five criteria, with only minor fluctuations. This stability suggests a more balanced and consistent application of the rubric across *Content*, *Organization*, *Vocabulary*, *Language Use*, and *Mechanics*.

From the analysis, many of the observed patterns do not exceed statistical thresholds significance. However, two comparisons approach significance. The first near-significant case appears in assessing *Vocabulary*. The target measure for the English Language Education group is 0.27, while the target measure for the 'Others' group is -0.40. The Rasch-Welch t statistic is 2.00, with a probability of 0.0511. Although slightly above the traditional 0.05 threshold, the result is not enough to suggest a meaningful trend. Therefore, this case implies no significant evidence to support the fact that raters from non-English programs are more lenient in assessing *Vocabulary* than those from English language education programs. Meanwhile, a second near-significant case appears within the 'Others' group, where their pattern shows noticeable fluctuation between scoring *Content* and *Vocabulary*. When scoring *Content*, their target measure is 0.28 logits, indicating moderate severity. When scoring *Vocabulary*, however, their severity drops to -0.53 logits, indicating leniency. The Rasch-Welch t-value for this contrast is 1.98 with a probability of 0.0517, which is not enough to prove significance.

Although most comparisons across groups do not reach statistical significance, the patterns in the line graph and target measures illustrate important trends. English Language Education raters remain stable, while English Literature and Language raters increase in severity on linguistically oriented criteria. The most striking pattern comes from the 'Others' group, which displays leniency toward *Vocabulary* despite being severe on *Content*. This finding is particularly interesting because in some English for Specific Purposes (ESP) contexts, vocabulary is a major instructional focus [19]. Yet in this dataset, raters in these contexts appeared lenient toward vocabulary accuracy.

One plausible explanation is that these raters prioritize content clarity and conceptual correctness, treating English primarily as a communicative tool for conveying information rather than as an object of linguistic evaluation. At the same time, the 'Others' group was the most severe on *Content*, which aligns with disciplinary norms emphasizing the correctness and completeness of ideas over linguistic detail.

5 Discussion

Based on the data from this study, several patterns of interaction effects were observed across raters' background groups. Although not all of these differences reached statistical significance, the graphical patterns clearly show variation in how different groups of raters applied the rubric criteria. The only significant interaction effect cases here occurred between raters' professional ranks and rubric criteria.

Based on interaction analysis between raters' gender and rubric criteria, in general, we can argue that male raters may prioritize macro-level features such as content quality and argument structure, while female raters may emphasize micro-level features such as language accuracy and mechanics. However, data analysis shows that these cases are not significant. This pattern is consistent with previous study that shows male and female raters do not assess differently in the context of EFL writing [20].

Previous research has suggested that raters with a background in language education differ from those with a background in linguistics or literature [21, 22]. Another earlier study has also reported differences between raters with linguistics majors and those with literature majors, with linguistics majors being more lenient [1]. In the Indonesian context, however, linguistics and literature are often housed within the same academic unit and may even belong to the same program of study, especially at the graduate level. For this reason, linguistics and literature were treated as a single category in this study. In contrast, language education represents a separate academic field with different training and assessment orientation. Therefore, the grouping used in this study reflects the local higher education context rather than distinctions commonly used in other research settings. In the present study, differences based on educational field were observed, particularly for the criteria of *Content*, *Language Use*, and *Mechanics*. In contrast, differences for *Organization* and *Vocabulary* were minimal and very small in magnitude, with bias contrasts close to zero. Although these differences were not statistically significant, the graphical patterns indicate that raters with language education background tended to be more severe on *Content*, while raters with a linguistics or literature background tended to be more severe on *Language Use*. These patterns suggest differences in focus rather than clear differences in overall severity.

Previous studies have reported that more experienced raters tend to be more severe in their scoring (e. g. [23–26]). However, those studies did not include analysis on lecturers' professional ranks, especially in higher education contexts. Previous study has also argued that less experienced raters tend to focus more on lexicogrammatical

features such as grammar, mechanics, and vocabulary, especially when evaluating weaker texts, while more experienced raters focus more on content and reasoning in higher quality texts [25]. While less experienced in rating or assessing might be related with entry-level rank in formal institutions, it cannot always be concluded that way. In our data, some participants have more than five years of teaching experience in universities, yet they still are in the 'Instructor' rank. Therefore, more studies are needed to clarify if raters with entry-level rank in higher education institutions shows more severity when rating vocabulary aspect of EFL writings.

Overall, the findings show that raters behave differently based on their background when assessing essays across rubric criteria, even though many of these differences are not statistically significant. The results suggest that background factors influence how raters interpret and apply rubric criteria. Thus, it leads to different patterns of severity and leniency at the criterion level. While these effects are not strong enough to produce consistent significant contrasts, they still point to meaningful variation in rater behavior that should be considered in writing assessment practice and rater training.

6 Conclusion

The analysis shows that demographic and professional background factors are associated with different patterns in how raters apply rubric criteria, although most of these patterns were not statistically significant. When examined graphically, however, the differences were clearly visible. In terms of gender, male lecturers tended to be more severe in assessing *Content* and *Organization*, while being more lenient in *Language Use*. In contrast, female lecturers tended to be more lenient in *Content* but more severe in *Vocabulary* and *Language Use*.

In addition to genders, differences were also observed based on educational qualifications. Lecturers holding doctoral degrees were more severe in *Content* and *Language Use*, but relatively more lenient in *Mechanics* and *Organization*. Moreover, differences were also found across academic fields. Lecturers with backgrounds in linguistics and English literature tended to be more severe in *Language Use* but more lenient in *Content*. The opposite pattern was observed among lecturers with backgrounds in English language education, who were more severe in *Content* and more lenient in *Language Use*. In addition, lecturers teaching in non-English programs or English for Specific Purposes contexts tended to prioritize *Content*. They were generally more severe in *Content* assessment and more lenient in *Vocabulary*. These patterns suggest that raters' academic and professional contexts shape how they prioritize different aspects of writing.

As shown in this study, lecturers at the early stages of their careers tend to show less stable rating patterns, despite being academically qualified. This suggests a need for targeted professional development. Institutions are encouraged to include assessment focused components in training programs for early career lecturers, particularly training

related to performance assessment and rater mediated assessment in the context of English as a foreign language. Beyond technical aspects of scoring, such training should also address rating behavior, including how to maintain consistency and avoid biases towards all rubric criteria.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. He, T.-H., Gou, W.J., Chien, Y.-C., Chen, I.-S.J., Chang, S.-M.: Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. Psychol Rep. 112, 469–485 (2013). https://doi.org/10.2466/03.11.PR0.112.2.469-485.
2. Erguvan, I.D., Dunya, B.A.: Analyzing rater severity in a freshman composition course using many facet Rasch measurement. Lang Test Asia. 10, 1 (2020). https://doi.org/10.1186/s40468-020-0098-3.
3. Yan, X., Chuang, P.-L.: How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. Language Testing. 40, 153–179 (2023). https://doi.org/10.1177/02655322221074913.
4. Wind, S.A., Peterson, M.E.: A systematic review of methods for evaluating rating quality in language assessment. Language Testing. 35, 161–192 (2018). https://doi.org/10.1177/0265532216686999.
5. Eckes, T.: Introduction to Many-Facet Rasch Measurement. Peter Lang D (2015). https://doi.org/10.3726/978-3-653-04844-5.
6. Eckes, T.: Rater types in writing performance assessments: A classification approach to rater variability. Language Testing. 25, 155–185 (2008). https://doi.org/10.1177/0265532207086780.
7. Ishikawa, S.: ICNALE: The International Corpus Network of Asian Learners of English, https://language.sakura.ne.jp/icnale.
8. Ishikawa, S.: The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. Routledge, London (2023). https://doi.org/10.4324/9781003252528.
9. Chang, Y.-T., Choe, A.T., Holden, D., Isbell, D.R.: Making each point count: Revising a local adaptation of the Jacobs et al.'s (1981) ESL COMPOSITION PROFILE rubric. Language Testing. 41, 443–455 (2024). https://doi.org/10.1177/02655322231217979.
10. Jacobs, H.L., Zingraft, S.A., Wormotuh, D.R., Hartfield, V.F., Hughey, J.B.: Testing ESL Composition: A Practical Approach. Newbury House, Rowley, MA (1981).
11. Alghizzi, T.M., Alshahrani, T.M.: Effects of grading rubrics on EFL learners' writing in an EMI setting. Heliyon. 10, e36394 (2024). https://doi.org/10.1016/j.heliyon.2024.e36394.
12. Heidari, N., Ghanbari, N., Abbasi, A.: Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. Lang Test Asia. 12, 20 (2022). https://doi.org/10.1186/s40468-022-00168-3.
13. Ishikawa, S.: The ICNALE Global Rating Archives: A new assessment dataset for learner corpus studies. Learner Corpus Studies in Asia and the World. 6, 13–38 (2024). https://doi.org/10.24546/0100487713.
14. Wind, S.A., Ge, Y.: Detecting rater biases in sparse rater-mediated assessment networks. Educational and Psychological Measurement. 81, 996–1022 (2021). https://doi.org/10.1177/0013164420988108.
15. Linacre, J.M.: Minifac64, (2025).

16. Han, C.: Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. Assessment & Evaluation in Higher Education. 43, 979–994 (2018). https://doi.org/10.1080/02602938.2018.1424799.
17. Lim, G.S.: The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. Language Testing. 28, 543–560 (2011). https://doi.org/10.1177/0265532211406422.
18. Lumley, T., McNamara, T.F.: Rater characteristics and rater bias: implications for training. Language Testing. 12, 54–71 (1995). https://doi.org/10.1177/026553229501200104.
19. Lu, F.-C., Chang, B.: Role-play game-enhanced english for a specific-purpose vocabulary-acquisition framework. Educational Technology & Society. 19, 367–377 (2016).
20. Bijani, H., Khabiri, M.: Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment : A FACETS analysis. The Journal of AsiaTEFL. 14, 687–702 (2017). https://doi.org/10.18823/asiatefl.2017.14.4.7.687.
21. He, T.: The impact of computers on marking behaviors and assessment: A Many-Facet Rasch Measurement analysis of essays by EFL college students. Sage Open. 9, 2158244019846692 (2019). https://doi.org/10.1177/2158244019846692.
22. Schoonen, R.: Generalizability of writing scores: An application of structural equation modeling. Language Testing. 22, 1–30 (2005). https://doi.org/10.1191/0265532205lt295oa.
23. Khamboonruang, A.: Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A Many-Facets Rasch Measurement approach. PASAA Journal. 66, (2023).
24. Mohd Noh, M.F., Mohd Matore, M.E.E.: Rating performance among raters of different experience through Multi-Facet Rasch Measurement (MFRM) model. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 11, 147–162 (2020). https://doi.org/10.21031/epod.662964.
25. Şahan, Ö., Razı, S.: Do experience and text quality matter for raters' decision-making behaviors? Language Testing. 37, 311–332 (2020). https://doi.org/10.1177/0265532219900228.
26. Tavassoli, K., Bashiri, L., Pourdana, N.: Experienced and novice L2 raters' cognitive processes while rating integrated and independent writing tasks. JLE. 8, 169–181 (2022). https://doi.org/10.17323/jle.2022.13466.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

