

Do student ratings match teacher ratings? Comparing rater severity and reliability in L2 writing assessment

Nurul Fariena Asli^{a,*}, Mohd Effendi @Ewan Mohd Matore^a, & Nurul Iman Ahmad Bukhari^b

^aThe National University of Malaysia (UKM), Malaysia

^bUniversity Malaysia Kelantan (UMK), Malaysia

Abstract

This study examines rater severity and rater reliability in a writing assessment across teacher and student raters within the Malaysian secondary schools L2 context. Four Primary Trait Writing (PTW) rubrics were developed and validated to be utilized in this self-assessment activity. A quantitative research approach was employed, involving 149 secondary school students and three English language teachers from six public schools in Malaysia. Teacher assessment and student self-assessment were used to evaluate students' essays, and the Many Facet Rasch Model (MFRM) was applied using FACETS to calibrate and analyse the scores. The analysis focused on identifying differences in rater severity and examining the consistency of rubric application between teacher and student raters. Results showed that teacher and student raters exhibited similar severity levels, with no statistically significant differences between the two groups ($\chi^2 = 0.0$, $df = 1$, $p = .90$). Furthermore, acceptable fit statistics ($MnSq = 0.5-1.5$; $Zstd$ within ± 2) indicated consistent application of the PTW rubrics across both rater groups. Overall, these findings suggest that both teacher and student raters shared a similar common understanding of the assessment criteria and functioned with comparable levels of as a homogeneous rater group, demonstrating similar levels of severity and reliability in the assessment of L2 writing performance. This indicates that students' self-assessment can align closely with teacher assessment when supported by structured rubrics, highlighting its potential value in L2 writing assessment contexts.

Keywords: Self-Assessment, Rater-Mediated Assessment, MFRM, Language Testing

Received: 17 June 2026

Revised: 25 September 2026

Published: 31 October 2026

1. Introduction

The educational assessment environment has evolved over the past decade from conventional teacher-centered evaluation toward more collaborative and learner-focused approaches. Contemporary frameworks increasingly position students as active participants in evaluating and managing their own learning rather than passive recipients of grades (Stranford, 2024). This shift reflects broader changes in educational philosophy that emphasize learner autonomy, reflective practice, and formative assessment as integral to instruction (Baidoo-Anu et al., 2023). In second language (L2) writing classrooms, these developments have supported the integration of self-assessment alongside teacher evaluation, reframing assessment as a pedagogical tool for learning rather than solely a mechanism for measurement and accountability (Yan, 2022).

In L2 writing teaching, self-assessment has garnered increasing academic interest due to its capacity to improve students' metacognitive awareness, self-regulation, and involvement in writing activities. Students are prompted to engage in self-evaluation to critically evaluate their writing skills, recognize strengths and limitations, and track their development based on established assessment criteria (Al-Nafjan et al., 2025; Khojasteh et al., 2025). These methods correspond with the ideas of formative assessment, which prioritize feedback, reflection, and ongoing enhancement above solely summative evaluation. In the Malaysian educational setting, enhanced theoretical and pedagogical support for self-assessment has prompted teachers to reevaluate the role of students in classroom evaluation and to encourage increased learner participation in assessment processes (Asli et al., 2024). Nonetheless, the effective execution of self-

* Corresponding author.

E-mail address: bm-5673@moe-dl.edu.my

assessment is significantly contingent upon the use of relevant and reliable assessment tools that may meaningfully and consistently inform students' evaluations.

Rubrics are among the most commonly utilized assessment techniques in writing evaluation, as they offer explicit performance standards and clearly defined descriptions for varying levels of achievement (Han et al., 2026). Rubrics enhance criterion-referenced evaluation by elucidating expectations and enabling more systematic scoring methodologies (Gonsalves & Lin, 2025; Hasan & Jones, 2024). In writing assessment, rubrics fulfill an educational role by aiding learners in comprehending the attributes of proficient writing and facilitating the alignment of their work with evaluation criteria. These attributes render rubrics especially appropriate for formative and self-assessment environments, where clarity and uniformity are paramount. Nonetheless, despite the widespread application of rubrics in educational contexts, a significant methodological concern remains regarding the equivalence of rubric functionality across diverse rater groups, including teachers and students.

The matter of rubric equivalence is crucial, as discrepancies in raters' interpretations and applications of assessment criteria may jeopardize the validity and reliability of assessment results (Hong, 2025). In rater-mediated assessment environments, scores are influenced not only by the quality of student performance but also by the subjective evaluations of raters. Educators and learners may significantly vary in their understanding of rubric descriptors, standards of writing quality, and acceptance of errors or deficiencies in written works (Alghizzi & Alshahrani, 2024). As a result, discrepancies in scoring rigor and uniformity may arise, possibly compromising the integrity of assessment outcomes. If students interpret rubric requirements differently from teachers, self-assessment scores may not correctly represent actual writing performance, therefore diminishing the instructional value of self-assessment activities.

The execution of rater-mediated assessment continues to be methodologically complex due to the intrinsic heterogeneity linked to human judgment. Studies in language testing and assessment have consistently shown that raters exhibit variability in their scoring behaviours due to differences in cognitive processing, interpretative frameworks, assessment experience, and personal biases (Lestari & Brunfaut, 2023). Certain raters may display more severity, continuously issuing lower scores, while others may show leniency by granting relatively higher ratings. Furthermore, raters may exhibit variability in their adherence to scoring guidelines, leading to inconsistencies in rating reliability. This variability constitutes a substantial source of measurement error and may undermine the fairness and accuracy of evaluation results. As a result, the examination of rater severity and reliability has emerged as a significant issue in the domain of language assessment and testing.

To mitigate these issues, sophisticated psychometric models have been progressively utilized to analyze rater behavior and assess the validity of evaluations. The Many Facet Rasch Model (MFRM) is an approach that enhances the Rasch measurement framework by integrating various dimensions of assessment procedures, such as examinee ability, rater severity, task difficulty, rating scale functionality, and assessment criteria (Cai & Yan, 2024). MFRM allows researchers to separate the several sources of variability affecting assessment outcomes, in contrast to typical scoring analyses that concentrate mostly on raw scores. The logit-based calibration offers comprehensive statistical insights into the relative severity and dependability of raters, the complexity of rating criteria, and the efficacy of rating scale categories. MFRM provides visual representations of interactions among facets, enhancing the interpretation of rating patterns in assessment contexts (Su & Shin, 2024).

The utilization of MFRM is especially beneficial in research examining self-assessment, as it facilitates systematic comparisons between teacher raters and student raters (Zheng et al., 2024). Through the analysis of rater severity and reliability, researchers can ascertain if students utilize rubric criteria in a manner analogous to those of teachers. If both groups exhibit comparable rating patterns and satisfactory dependability, this may offer empirical validation for employing self-assessment in classroom writing evaluation. Conversely, significant disparities between teacher and student evaluations may signify issues pertaining to rubric interpretation, assessment literacy, or rater training (Zheng et al., 2026). MFRM provides a comprehensive analytical methodology for validating the application of rubrics in educational contexts with diverse rater groupings.

In the context of Malaysian secondary schools, research on the comparability of teacher and student evaluations in L2 writing assessment is still rather scarce. Despite the growing advocacy for self-assessment in educational practice, scientific evidence concerning the efficacy of writing rubrics among various user groups remains inadequate. Previous research has predominantly concentrated on instructor evaluations or the educational advantages of self-assessment, neglecting a comprehensive examination of the psychometric characteristics of the utilized assessment tools. Moreover, little focus has been directed on the application of Primary Trait Writing (PTW) rubrics particularly formulated to evaluate particular writing attributes pertinent to classroom instruction. Consequently, additional inquiry is required to

ascertain the reliability and consistency of PTW rubrics when employed by both teachers and learners in Malaysian L2 writing classes.

This study seeks to assess the rater severity and reliability of teacher and student raters in the Malaysian secondary school L2 context using Primary Trait Writing (PTW) rubrics in assessing students' essays. To accomplish this goal, the Many Facet Rasch Model (MFRM) was utilized to analyze rating data and identify potential differences scoring between teacher and student raters. The study investigates the following research question: To what extent do teacher and student raters differ in rater severity and reliability when assessing L2 writing using PTW rubrics?

2. Literature Review

2.1. Rater-Mediated Assessment in Writing

Rater-mediated assessments are used extensively in a variety of educational contexts and are defined as assessments that involve raters assessing test-taker performance and employing rating scale categories to delineate performance levels across one or more areas (Engelhard & Wind, 2019; Su & Shin, 2024). Since writing performance cannot be measured directly, human judgment is an integral part of this assessment, even though the process is guided by a set of rubrics or marking criteria, as the interpretation might differ from one person to another. Due to that, the evaluations of the reliability, validity and fairness of these procedures involving the facets (a) examinees, (b) raters, (c) scoring criteria and (d) task, need to be considered in a rater-mediated assessment where the influence of these facets on the scores is unavoidable (Tanaka, 2025). As for writing tasks, raters are required to interpret complex constructs such as content development, organisation, and language use, which introduces variability in scoring (White et al., 2025). This has led to rater effects, such as severity, leniency, inconsistency, and bias, that may threaten the reliability and fairness of writing scores if they are not adequately examined and controlled (Neittaanmäki & Lamprianou, 2024). In the classroom context, the unwanted variability becomes more complex when self-assessment is conducted, as it involves not only the teachers but the students as well. Differences in assessment literacy and interpretation of rubric descriptors, which influence the scoring behaviour, may occur if a suitable tool is not used to address the issues.

Prior studies on rater-mediated assessments in language contexts generally focus on teachers, language experts, and graduate-level participants, leaving an evident gap in understanding the role of students as raters evident in the previous studies by (Ahmadi Shirazi, 2019; Humphry & Heldsinger, 2019; Lertsakulbunlue & Kantiwong, 2024; Mohd Noh & Mohd Matore, 2020; Mohd Noh & Mohd Matore, 2022; Şahan & Razi, 2020). However, one of the few studies on self-assessment where students acted as raters has shown that student ratings often correlate moderately with teacher ratings (Chang & Tseng, 2025). Teacher raters tend to apply more consistently due to their assessment experience, in contrast with student raters, who may demonstrate greater variability and leniency. More recent studies using rater-mediated analyses, such as Multi-Faceted Rasch Measurement, have further highlighted systematic differences in severity and bias between rater types, suggesting that rater background influences how assessment criteria are interpreted and applied (Han et al., 2026; Li et al., 2025). These findings indicate the need to empirically examine raters when alternative rater groups, such as students, are involved in assessment, particularly when rubrics are used to support authentic classroom-based evaluation. Rater severity denotes the inclination of certain raters to consistently issue higher or lower ratings relative to their peers assessing the same performance. Strict raters typically provide lower scores, while tolerant raters generally confer better grades. Inconsistency arises when raters apply scoring standards erratically across performances, whereas bias occurs when extraneous variables affect rating decisions. This variability jeopardizes the fairness, reliability, and validity of assessment results, especially in classroom-based evaluations with numerous raters.

2.2. The use of Many-Facet Rasch Measurement (MFRM) in Determining Severity and Reliability

MFRM was selected due to its capacity to observe and calibrate discrepancies in rater severity, ensuring that variances across raters are considered in score interpretation (Cai & Yan, 2024; Y. Li et al., 2025). MFRM, unlike conventional analysis methods, does not require uniform ratings from raters; it explicitly accounts for variations in severity and consistency. This study employs MFRM to systematically examine discrepancies in student and instructor evaluations, thereby demonstrating the reliability and equity of the rubrics among various raters. This method facilitates the

modelling of each item and judge based on the use of the rating scale, and it may statistically assess interactions among various facets.

The basic MFRM model is expressed as:

$$\log \frac{P_{nijk}}{P_{nijk-1}} = B_n - D_i - C_j - F_k$$

where P_{nijk} is the probability of examinee n receiving a rating k on item i from judge j ; B_n is the examinee's ability; D_i is item difficulty; C_j is judge severity; and F_k represents the additional difficulty associated with category k relative to $k - 1$ (Cai & Yan, 2024).

Many-facet Rasch measurement (MFRM) has become the leading framework for measuring rater severity and rater reliability. This approach treats raters as one facet of measurement alongside students and tasks, allowing estimation of each rater's severity on a common logit scale (Uto et al., 2024). MFRM assigns severity and leniency logit measures to raters, which indicate how strictly or leniently each rater's scores are compared to the overall rating scale for the same performance (Huang & Chen, 2022). Raters with positive logit values are more severe (harsh), while those with negative values are more lenient. If significant differences in severity are observed among raters, this could suggest subjectivity in rubric interpretation, which may require additional sessions to ensure consistency. In terms of writing assessment, the severity of the rater is considered a characteristic that can be adjusted and modelled using rubrics and advanced statistical methods primarily for score interpretation (Azizah et al., 2020). Since this study is based on two groups of raters, the logits are based on rater type measurement, but apply the same value as explained. The FACETS map, depicted in Figure 2 and sometimes known as the vertical ruler table or Wright map, provides a visual representation that establishes a structured framework for understanding the relevance of scores derived from the evaluation. It offers visual data on discrepancies that may occur among different elements of a facet, including variances in the severity and expertise of raters between writers. The estimation of writer or examinee ability logit measures occurs simultaneously with the estimation of rater severity logit and assessment criterion difficulty logit. Aligning them on a uniform linear measurement scale facilitates straightforward comparison of the results.

Apart from that, raters must be able to use the rubrics in their ratings consistently across different levels of student proficiency (Panadero et al., 2023; Panadero & Jonsson, 2020). MFRM provides several reliability indicators to assess this consistency. In addition to separation reliability, infit and outfit mean square statistics help assess whether raters apply the rubric as intended (Asli et al., 2024). The infit statistic, which is information-weighted, detects inconsistencies in a rater's scoring pattern, while the outfit statistic identifies extreme scoring behaviours. Acceptable values generally range between 0.5 and 1.5, indicating that raters are using the rubric appropriately. If a rater exhibits misfit (values exceeding 1.5), it may suggest difficulties in interpreting the rubric, necessitating further rater training or revisions to the rubric descriptors. One key statistic is rater separation reliability, which measures the extent to which the model can distinguish more severe raters from more lenient ones (Ariakpomu et al., 2025). A high separation reliability value (typically ≥ 0.80) suggests that raters are consistently different in severity but stable in their rating behaviour, indicating the rubric allows for meaningful differentiation. Conversely, a low value may imply that raters are scoring too similarly, which could mean that the rubric does not provide sufficient detail for effective differentiation. However, from a different perspective, in rater-mediated assessment, a degree of rater unison or shared understanding of the scoring criteria is desirable, as it suggests consistent application of the rubric across raters (Palemo, 2022). Within the MFRM framework, rater separation reliability reflects the extent to which differences in rater severity can be detected. Therefore, lower separation may indicate more homogeneous rater behaviour, which can be interpreted as closer alignment in scoring tendencies, provided that overall fit statistics and measurement quality are maintained.

Recent studies using MFRM have highlighted severity and reliability differences across groups such as native vs non-native speakers, novice vs experienced raters, and classroom teachers in L2 writing assessment. A study by Esfandiari & Myford (2020) compared the severity differences between 6 teachers and 188 students in two Iranian universities, where self-raters showed more variations in their severity level and teacher raters showed more severe ratings. Recently, Khamboonruang (2023a) conducted a study on 53 EFL Thai undergraduates' rating performances in an ongoing classroom assessment, which revealed a consistent pattern where self-assessors are more lenient and heterogeneous compared to peer-assessors and teacher-assessors, who are more severe. Self-assessment showed more inconsistencies compared to peer assessment; however, both groups are able to distinguish levels of writing quality, which proves that their ratings are not random. Both studies showed the same results, where overall, self-assessment tends to be more lenient compared to teacher rating, even though rater training was conducted before the scoring process. Meanwhile, another study by Khamboonruang (2023b) investigated experienced and inexperienced teachers' ratings of opinion

essays written by 42 English-major undergraduates. The results found that teachers were not uniform in their severity levels; however, the three teachers were consistent in how they applied the rating scale when evaluating student essays. The results from those studies were gained based on the scores from the raters while using an analytic type of scoring scale. Thus, the question arises of how to minimise the unwanted variability in interpreting the scales. It is often compromised by factors such as varying interpretations of rubrics, differing weighting of assessment criteria, and individual rater biases (Algeciras & Oliden, 2023), primarily within the context of self-assessment in writing, as it is very complex to establish consistent and accurate evaluations.

2.3. *Self-Assessment*

Modern educational theories increasingly prioritize learner-centered methodologies, wherein students actively engage in monitoring, regulating, and assessing their own learning advancement (Yan, 2022). From this viewpoint, self-evaluation is regarded as a crucial element of assessment for learning, as it stimulates learners to engage in critical reflection on their performance and fosters more accountability for their educational results. Self-assessment is conceptually linked to constructivist learning principles, wherein knowledge is actively developed through reflection and engagement with learning tasks. Students utilize self-evaluation procedures to engage in metacognitive processes, enabling them to analyze their strengths and deficiencies according to defined assessment criteria. This technique fosters the advancement of self-regulated learning, assessment literacy, and learner autonomy. (Yan, 2022) asserts that self-evaluation allows students to internalize assessment criteria and synchronize their learning behaviors with educational objectives. In language learning environments, especially in L2 writing, self-assessment enhances awareness of linguistic precision, coherence, organization, and communicative efficacy. The efficacy of self-evaluation is largely contingent upon the clarity and quality of the employed assessment criteria. In the absence of explicit direction, students may find it challenging to assess their own work effectively, leading to inconsistent or biased evaluations. Consequently, the implementation of meticulously crafted rubrics is vital for facilitating substantive self-assessment activities. In this study, the PTW rubrics serve as mediating instruments that systematically assist both teacher and student raters in assessing writing performance. The alignment between teacher assessments and self-assessments is thus regarded as evidence that students can comprehend and effectively utilize the rubric criteria.

2.4. *The use of rubrics in writing assessments*

Another crucial aspect in this study pertains to rubric-based assessment in writing assessment. The evaluation of writing is intrinsically intricate, as writing proficiency cannot be directly quantified by objective testing methods. Evaluation mostly relies on human judgment, rendering scores vulnerable to subjectivity and variability (El-Tallawy et al., 2024). Rubrics are extensively utilized to mitigate score inconsistencies by offering clear criteria and performance descriptors that standardize grading processes. Rubrics are typically defined as scoring instruments that delineate different performance levels based on established criteria (Lertsakulbunlue & Kantiwong, 2024). In writing assessment, rubrics help raters evaluate multiple dimensions of writing quality, including content, organisation, language use, grammar, vocabulary, and mechanics. In addition to their evaluating role, rubrics fulfill instructional functions by elucidating expectations and enhancing students' comprehension of quality performance.

This study employs the Primary Trait Writing (PTW) rubric, which is in contrast to holistic rubrics, as this type of rubrics yield a singular overall score, and analytic rubrics, which assess numerous general criteria concurrently, PTW rubrics concentrate on specific writing attributes pertinent to the objectives of a designated writing job (Roberts & Whitfield, 2025). This trait-centric methodology enables evaluators to focus on distinctly articulated performance factors, thus reducing interpretative uncertainty. In L2 writing environments, PTW rubrics are advantageous as they offer explicit and targeted instruction that aids both educators and students in achieving more consistent writing evaluations (Yu et al., 2025). The theoretical efficacy of rubrics hinges on the consistent interpretation and application of criteria by various raters. When teachers and learners possess a mutual comprehension of rubric descriptors, the inconsistency in scoring can be reduced, thereby enhancing the reliability and validity of assessments. Inconsistent interpretation of rubric requirements may lead to variations in severity, leniency effects, or rater bias.

In self-assessment environments, these issues are exacerbated due to students typically having less assessment experience than teachers. Students may variably interpret rubric descriptors, inconsistently apply criteria, or exhibit leniency in self-evaluation. Thus, analyzing rater behavior is essential in ascertaining the reliability of self-assessment as an evaluative technique in L2 writing classes. This study examines the disparities between instructor raters and student raters on severity and reliability in the use of PTW rubrics. The theoretical premise underlying this comparison

posits that if the rubrics are adequately explicit and accompanied by suitable training, both groups should exhibit comparably similar scoring behaviors.

3. Methods

3.1. Research Design

This study utilized a quantitative research approach to evaluate the severity and reliability of teacher and student raters in utilizing the PTW rubrics to assess students' writing performances. This included classroom-based self-assessment procedures by designating students as active raters of their own writing. This design adheres to formative assessment concepts that prioritize student engagement, reflection, and assessment literacy. The use of PTW rubrics is believed to facilitate uniform interpretation of assessment criteria among various rater groups by comparing teacher assessment scores with student self-assessments. The comprehensive research flow is depicted in Figure 1.

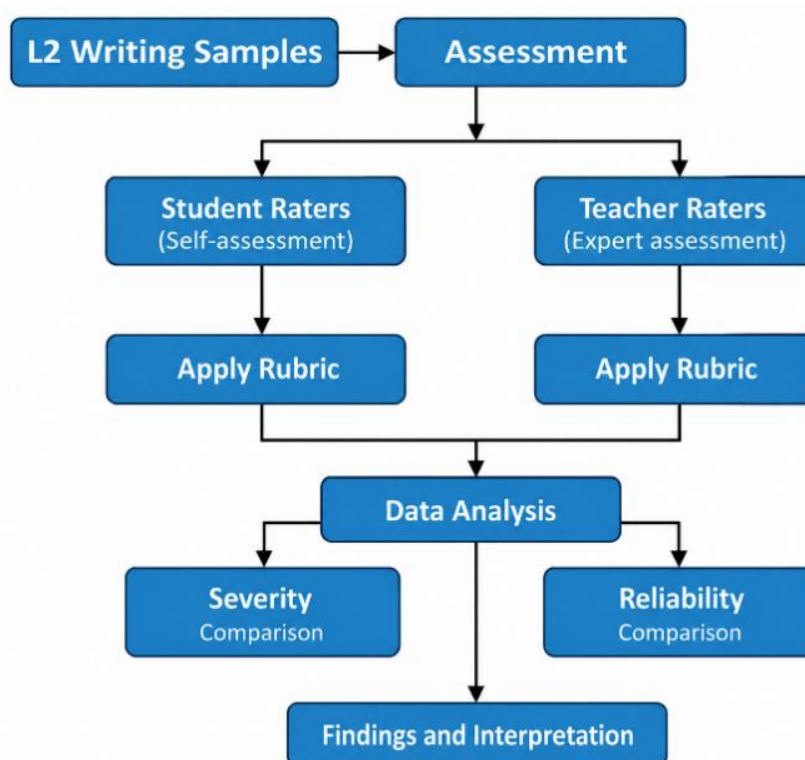


Figure 1. Research flow.

3.2. Design and Development of the PTW Rubrics

The PTW rubrics employed in this study were particularly formulated to evaluate L2 writing performance within the Malaysian secondary school framework. The rubric construction process was informed by the ideas of Primary Trait Writing assessment, emphasizing the evaluation of specific writing characteristics pertinent to the objectives of a designated writing task. In contrast to holistic and analytic rubrics, the PTW approach offers more precise and task-oriented descriptors that enhance clarity of interpretation for both teachers and students. The development approach commenced with the identification of writing structures pertinent to CEFR-aligned classroom writing activities. Four fundamental writing features were chosen based on curriculum mandates and writing goals: topic relevancy, communicative effectiveness, coherent devices, and format. Each feature was operationalized into performance

descriptors that delineate various levels of writing proficiency. The descriptors were meticulously crafted to guarantee clarity, precision, and accessibility for secondary school students participating in self-assessment activities.

The creation of the rubric included recurrent refinement processes to guarantee consistency between theoretical notions and practical classroom implementation. The creation process involved a comprehensive examination of existing literature on rubric development, classroom assessment, and L2 writing evaluation. Special emphasis was placed on ensuring that the descriptions were sufficiently unambiguous to minimize interpretative ambiguity and diminish undesirable rater variability. Subsequent to the original development phase, the PTW rubrics were subjected to expert evaluation and preliminary validation processes before their application in the current study. The rubrics were subsequently implemented in real classroom assessment environments with both teacher and student evaluators. The empirical validation approach aimed to assess whether the rubric might facilitate uniform scoring behaviors between the two groups. The validity and reliability evidence was later assessed by MFRM analysis.

3.3. Participants

A total of 149 Form 4 students and three English teachers from six national public schools in Malaysia participated in this study. The participants were drawn exclusively from national secondary schools to ensure a relatively homogeneous educational background, as all students were assumed to have received comparable instruction and exposure to the same syllabus. A purposive sampling technique was employed to select participants based on three levels of English writing proficiency represented by 6 performance standards (TP). To represent the high-proficiency group, 50 students who achieved TP5 and TP6 in English writing were selected. The intermediate group consisted of 50 students who obtained TP3 and TP4, while the lower-proficiency group comprised 49 students with TP1 and TP2 in writing. This stratification enabled a balanced representation of learners across proficiency levels for subsequent analysis.

3.4. Data Collection Instruments and Validation Procedures

Ensuring connectivity among all facets is essential in MFRM analysis. To achieve this, the study employed a Two-Facet Linked Design using anchor raters (Engelhard & Wind, 2019a), where expert teachers served as common reference points linking students, raters, and criteria. This approach allowed the MFRM model to account for differences in rater severity, maintain comparable scores, reduce bias, and strengthen the validity of the ratings (Asli et al., 2024). Before the actual assessment, students received detailed instruction on the PTW rubrics and were given opportunities to ask questions about unclear aspects. A sample essay was then rated together with the researcher, allowing students to observe how scores were applied and understand the expected features for each performance level. This rater training enabled novice student raters to calibrate their judgments with those of expert raters, ensuring more consistent and reliable scores.

3.5. Data Analysis

This study employed FACETS Version 3.67.1 software to assess the severity and reliability of teacher and student raters in using PTW rubrics.

$$\log \frac{P_{nijk}}{P_{nijk-1}} = B_n - D_i - C_j - F_k$$

For this study, the above MFRM equation shows that (B_n) represents the writing ability of student (n) , which indicates the underlying quality of the essay being assessed. (D_i) represents rater severity, referring to the degree to which a rater tends to award higher or lower scores compared to other raters. (C_j) represents the difficulty of trait (j) in the PTW rubric, reflecting how challenging it is for students to achieve higher scores on a particular writing trait. (F_k) represents the threshold between adjacent rating scale categories, which defines the difficulty of moving from one score category to the next. In practical terms, the model separates observed scores into student ability and systematic effects of rater behaviour, trait difficulty, and rating scale structure, allowing more accurate interpretation of writing performance. Following calibration, model-data fit was examined to determine whether the observed ratings were consistent with the assumption that the assessment measured the intended latent writing construct. FACETS data presents a Wright map and a measurement report to provide the evidence that will answer the research question.

Apart from that, another crucial part of using MFRM was confirming that there was connectivity or linkage between all the facets that were calibrated. The researcher developed the two-Facet Linked Design judging plan to avoid a

disjointed link between facets as suggested by (Engelhard & Wind, 2019). A total of 149 student essays were used in the rubric implementation phase. For self-assessment, each essay was rated across four traits, resulting in 596 self-assessment scores (149 essays $\times$ 4 traits). In addition, three expert raters evaluated the same set of essays using the same four PTW traits, generating 1,788 rater-mediated scores (3 raters $\times$ 149 essays $\times$ 4 traits). This produced a combined dataset of 2,384 observable scores, which were subsequently analysed using the FACETS program.

4. Result and Discussion

4.1. Section 1: Unidimensionality Result

Prior to assessing the severity and reliability of teacher and student raters, it was essential to ascertain whether the instrument met the unidimensionality assumption within the Rasch measurement framework. Unidimensionality denotes the degree to which an instrument assesses a singular underlying notion instead of numerous unrelated dimensions. Establishing unidimensionality is crucial, as the validity of Rasch measurement relies on the premise that all rubric criteria contribute to the assessment of a shared latent characteristic. Table 1 displays the summary data derived from the Rasch dimension analysis conducted using FACETS software. The basic principle of the Rasch Model is unidimensionality, which focuses on one attribute or dimension at a time (Bond & Fox, 2001), and it must achieve more than 40% variance in the Rasch Model (Engelhard & Wind, 2019). The results demonstrate that the variance accounted for by the Rasch measures was 68.25%, surpassing the minimum acceptable level of 40% recommended in Rasch measurement literature (Royal, 2014). This outcome indicates strong support for the assumption that the PTW rubrics effectively assessed a predominant singular construct and that the evaluation data conformed to the Rasch model suitably. Simultaneously, the variance of residuals was 31.75%, signifying that some unexplained variance remained beyond the primary dimension. Although the percentage of variance explained by Rasch measures should not be interpreted as the sole indicator of dimensionality, these findings provide sufficient evidence to support the assumption of unidimensionality, these findings provide sufficient evidence to support the assumption of unidimensionality and the suitability of the PTW rubrics for subsequent Many-Facet Rasch Measurement (MFRM) analyses of rater severity and reliability.

Table 1. Unidimensionality.

Statistics	Count	Mean	S.D.	Parameters
Responses used for estimation	2352	2.74	1.23	297
Responses in one extreme score	32	4.00	0.00	8
All responses	2384	2.76	1.23	305
Count of measurable responses	2384			
Raw-score variance of observations	1.51	100.00%		
Variance explained by Rasch measures	1.03	68.25%		
Variance of residuals	0.48	31.75%		

4.2. Section 2: Severity Level, Rater Separation, and Reliability of Self-Assessment and Teacher Assessment

Primarily, the Wright map (Figure 2) is presented to show the location of the rater type facet: Self-Assessment (SA) and Teacher Assessment (TA). The Wright map serves as a visual representation indicating the position of facets along the continuum that characterises a latent variable or construct. The polytomous Rasch Model (MFRM) enables the acquisition of invariant parameters when a good model-data fit is achieved. The parameters will provide significant and practical measures that can enhance the Wright map. Estimation is the statistical method for identifying individuals and indicators (domain, items, and thresholds) considered as unknown quantities. The unknown quantities in measurement are the parameters within the measurement models. The estimation parameter for all facets in MFRM will be calibrated according to the model's data fit. In other terms, it is referred to as the fit statistic.

As for the Wright map, the logit scale on the left column is the same for all facets. It will find the logit value for each facet. The names of the aspects used in this investigation are at the top of the map. The titles of the facets also have “+” and “-” signs next to them to show if a facet metric is positive or negative. For example, the examinee (facet 1) is measured positively, indicating that the higher the students' position in the Wright Map, the higher the students' ability. On the other hand, raters are rated negatively, so the higher the rater's location on the map, the more severe the rater is. FACETS also measured an item or trait negatively, which means that the higher the Rasch measures, the harder the characteristic is. Elucidated this as a convention in educational assessment, whereby establishing the benchmark for

data calibration (Jin & Eckes, 2026). The second column has a number from 1 to 149 representing the students who are participants in this study. The third column shows the raters' numbers, which go from 1 to 152. Raters 1 to 149 are students, whereas raters 150 to 152 are teachers. The fourth column shows where the rater type is located, and it looks like both types are in the same location. The fifth column presents the traits in order, with the hardest attribute at the top and the simplest of the four traits at the bottom. "Format," which is at the bottom, is the easier criterion for giving a higher score. On the other hand, "Cohesive Devices" is the criterion that is harder for the raters to give a high grade. Most items are clustered in the middle of the distribution, between -1.00 and 1.00 logit. The dispersion of all the criteria shows that raters could tell the difference between them or that they are not related. Lastly, the fifth column shows the PTW rubrics' four-point rating scale. There are five categories, from 0 to 4, with 0 being the lowest and 4 being the highest.

Figure 2 shows that the locations of SA and TA in the Wright map are at the same level, which means that SA and TA are at the same level of severity. Apart from that, the measures to determine the severity level of both rater types were taken from Table 2. The severity of the raters in rating behaviour for both teacher and student raters was evenly distributed around 0.00 logit, as indicated in column Measure of Table 2. The measured value is organised in ascending order of severity, ranging from the most severe to the most lenient. The ranking becomes more severe as the measure increases and more lenient as the measure decreases (Li & Wang, 2021; Li, 2022). The measure of student raters' severity was 0.01 logits, but the measure of teacher raters' severity was 0.00. Evidence indicates that the scores given by student raters were slightly more severe than the scores given by teachers, and it can be suggested that there is no significant difference between students' self-ratings and their teachers' ratings.

Table 2. Rater Type Measurement Report.

Total Score	Obs Ave	Fair M-Ave	Mea	Infit Mnsq	ZStd	Outfit Mnsq	Zstd	Corr.	Total Score
1631	2.7	3.0	.01	.94	-.9	.96	-.5	0.55	Self-Assessment
4938	2.8	2.99	.00	.94	-1.5	.99	.0	0.44	Teacher Assessment

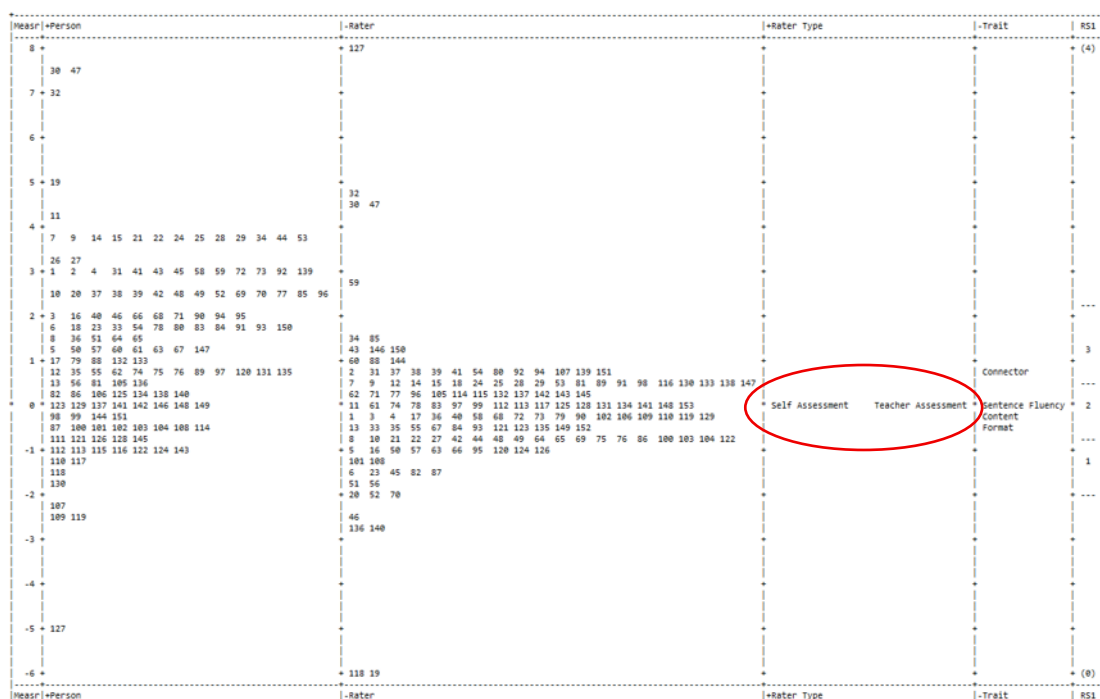


Figure 2. The Wright Map for rater type facets

4.3. Section 3: Consistency between Self-Assessment and Teacher Assessment

To assess the consistency of ratings between the two rater types, fit statistics were examined. As shown in Table 2, the self-assessment (SA) group had an infit MnSq of 0.96 (ZStd = -0.9) and an outfit MnSq of 0.96 (ZStd = -0.5). For the teacher assessment (TA) group, the infit MnSq was 0.94 (ZStd = -1.5), and the outfit MnSq was 0.99 (ZStd = 0.00). All infit and outfit Mean-Square (MnSq) values fell within the acceptable range of 0.5 to 1.5, indicating that both students' and teachers' raters demonstrated good internal consistency in their ratings. The standardised fit statistics (ZStd) were also within the commonly accepted limits of -2 to +2, further confirming that the ratings from the two rater groups were reliable, accurate and free from extreme misfit. These results suggest that the PTW rubrics were interpreted and applied consistently across different types of raters.

Table 3. Rater type discrimination measures

Measures	MFRM Output
Trait separation ratio	0.00
Trait separation index	0.33 (1 strata)
Chi-square	0.0, with 1 degree of freedom, p = 0.90
Reliability	.00

Based on Table 3, the chi-square analysis indicates that there was no statistically significant difference in the severity between the two categories of raters (chi-square = 0.0, df = 1; p = 0.90). The separation ratio was 0.00, indicating that there was no distinction between different levels of severity among the rater types. The strata index was 0.33, reflecting a moderate level of variability in the severity ratings. However, the reliability was 0.00, indicating that the severity ratings cannot be reliably divided into multiple levels. Due to a reliability score of 0, there is no possibility to distinguish the severity between the two rater types.

4.4. Discussion

4.4.1. Severity and Reliability Between Teacher and Student Raters

The results of this study provide evidence that students are capable of self-assessing their L2 writing with a level of severity and consistency comparable to teacher raters. Both teachers and students demonstrated similar patterns of severity and consistency as can be observed in the Wright Map and the rater measurement reports generated through the MFRM analysis. Both teacher assessments (TA) and self-assessments (SA) mapped onto almost identical points on the logit scale. The difference between the mean severity of the two rater groups was minimal (0.01 logits), suggesting a strong level of agreement in rater behaviour. While student raters were marginally more severe than teacher raters, the difference was practically negligible and well within acceptable educational ranges. This suggests that, when supported by a well-constructed PTW rubric, students are able to interpret and apply the assessment criteria in a manner comparable to trained teachers.

The separation reliability statistics further support this pattern of alignment, although their interpretation differs within the MFRM framework. Unlike traditional test reliability indices, rater separation reliability does not function as an indicator of measurement quality in a linear sense. Rather, a low separation reliability reflects minimal dispersion in rater severity estimates, indicating that raters operated at broadly similar levels of severity and were therefore not strongly distinguishable by the model. A rater separation reliability of 0.00 indicates that the variability in rater severity was negligible and insufficient for the model to reliably distinguish teacher raters from student raters. In other words, the system cannot reliably separate teachers and students into different groups of grading severity. This is further supported by the stratum index of 0.33. A stratum index below 1.00 indicates very limited separation, suggesting that the model does not identify distinct severity tiers among raters. Rather than acting as “lenient students” and “harsh teachers,” both cohorts functioned as a relatively homogeneous grading group. Consistent with Jin and Eckes (2026), near-zero separation reliability may be interpreted as evidence of rater equivalence among rater groups rather than measurement failure. Overall, the findings suggest that the PTW rubric promoted comparable severity between teacher and student raters, with both groups applying similar evaluation thresholds.

These outcomes offer a refreshing contrast to previous MFRM research in L2 writing. Past studies frequently indicated that teacher raters are notably harsher than student self-assessors, and that self-assessment tends to exhibit significantly greater variability. However, much of this previous research relied heavily on highly detailed, multi-dimensional analytic rubrics. While analytic rubrics are excellent for diagnostic feedback, they often impose a heavy cognitive load on non-specialist student raters, which frequently leads to the leniency and high score variance observed in prior

literature. By showing no significant gap in severity, our study suggests that the specific structural design of the PTW rubric may have streamlined this assessment process, successfully mitigating the undesirable rater variability and over-lenient biases that traditionally plague classroom-based self-assessment scenarios. Apart from that, previous research has often shown that students tend to rate their own writing more leniently than teachers, largely because of limited assessment literacy, limited familiarity with the criteria, and self-related or motivational factors (Panadero et al., 2023; Panadero & Jonsson, 2020). However, this tendency is not universal. Several scholars have argued that the accuracy of self-assessment depends on the quality of the assessment rubric, the clarity of its descriptors, and the extent of training students receive before evaluating their own work (Tian & Tong, 2026; Vasileiadou & Karadimitriou, 2021). When students are equipped with well-defined, criterion-referenced rubrics and receive adequate guidance on applying the assessment criteria, they are more likely to produce ratings that closely align with those of teachers. These findings suggest that differences between teacher and student ratings may reflect variations in assessment design and implementation rather than an inherent tendency for students to assess themselves leniently.

4.4.2. Role of Primary Trait Writing Rubrics in Minimising Rater Variability

The notable consistency between teacher and student raters can be primarily attributed to the design features of the Primary Trait Writing (PTW) rubrics. The PTW rubric employed in this study was deliberately designed to offer clear, targeted, and task-specific descriptions that directed raters towards a common comprehension of performance standards. In contrast to comprehensive analytic rubrics, primary trait rubrics focus exclusively on the fundamental writing characteristics pertinent to a certain assignment, thus minimizing interpretative ambiguity. Studies investigating the relationship between raters and scoring rubrics have constantly highlighted that the design of rating scales significantly affects assessment quality and scoring reliability (Uto et al., 2024). Rubrics with ambiguous or excessively general descriptors may permit raters to interpret criteria variably, resulting in increased inconsistency in scoring behavior. Conversely, PTW rubrics offer succinct and precise delineations of anticipated performance, allowing evaluators to assess writing more methodically and uniformly.

The current findings corroborate previous assertions that primary trait scoring provides targeted and effective direction for writing evaluation (Asli et al., 2024). Prior research indicates that educators favor main attribute rubrics in classroom environments since they enable more precise and uncomplicated assessment of student writing (Veloo et al., 2018). In L2 writing teaching, such specificity is crucial as second language learners frequently need explicit direction to comprehend evaluation criteria and assess their own writing effectively (Safari & Ahmadi, 2025). Recent meta-analytic research indicates that rubric-based self-assessment enhances academic achievement, self-regulated learning, and learner self-efficacy across many educational contexts (Panadero & Jonsson, 2020). The results of this study enhance the existing literature by offering empirical evidence that meticulously crafted PTW rubrics can facilitate dependable self-assessment practices in rater-mediated L2 writing evaluation. The analogous severity and dependability noted across instructor and student evaluators suggest that the rubric effectively enabled a unified understanding of writing quality and performance criteria.

4.4.3. Contribution of Structured Rater Training

A significant aspect contributing to the consistency seen in this study was the execution of systematic rater training prior to the scoring process. Due to the relatively limited expertise of student raters in assessment methods, rigorous training was essential to acquaint them with the rubric criteria and scoring procedures. The training sessions included supervised practice scoring, collaborative dialogues, and the utilization of benchmark essays to create a shared comprehension of performance standards. The efficacy of rater training in enhancing scoring consistency has been extensively recorded in prior studies. Research indicates that training can mitigate rater severity and bias, enhance the scoring capabilities of rookie raters (Mohd Noh & Mohd Matore, 2020; Mohd Noh & Mohd Matore, 2022), and improve rating consistency among seasoned evaluators (Knoch et al., 2021). While training cannot entirely eradicate human variability in scoring behavior, it is essential for reducing unintended discrepancies and enhancing a collective understanding of evaluation standards.

The training protocols employed in this study adhered to evidence-based techniques advocated in writing assessment literature. Students engaged in structured scoring exercises utilizing sample essays and participated in discussion-based moderation sessions, enabling them to compare their evaluations with those of expert raters. These activities facilitated student raters in internalizing the rubric descriptors and refining their scoring selections with more precision. As a result, students enhanced their ability to assess their work against criteria similar to those utilized by teachers. Longitudinal studies of L2 writing assessment have underscored the significance of continuous rater training, as Wu et al. (2023) show that ongoing moderation, retraining, and feedback are essential to mitigate rater drift and uphold score

reliability over time. The current findings reinforce the notion that systematic training protocols are crucial for the implementation of self-assessment practices involving inexperienced student raters.

4.4.4. Implications for Self-Assessment in L2 Writing Classrooms

The results of this study have significant implications for classroom assessment procedures in L2 writing education. Previous MFRM research has sometimes shown discrepancies and leniency in self-assessment techniques, raising questions about the trustworthiness of scores produced by students. Nonetheless, the current data indicate that such fluctuation is not unavoidable. When students get well-structured rubrics and systematic rater training, they can perform self-assessments with a consistency comparable to that of teacher raters. These findings directly confront issues in the literature pertaining to conflicting interpretations of rating scales and rubric descriptors (Algeciras & Oliden, 2023). They underscore the essential need for specific rubric design in facilitating valid and reliable classroom-based writing evaluation. The study illustrates that self-evaluation should not be regarded solely as a pedagogical tool for enhancing learner engagement, but also as a potentially valid assessment method when underpinned by suitable measurement tools and training protocols. The findings underscore the significance of self-evaluation in fostering learner autonomy and assessment literacy from a comprehensive educational standpoint. By actively engaging in the evaluation of their own work, students enhance their awareness of assessment criteria and improve their ability to identify areas needing enhancement. Thus, self-assessment via PTW rubrics may enhance both dependable classroom assessment methodologies and the cultivation of thoughtful, self-regulated learners in L2 writing environments.

Other implications of these findings are in terms of educational policy and curriculum design. From a policy perspective, students are proven able to be responsible for their own learning; for example, practicing self-assessment with an appropriate level of severity and consistency, and the ratings are nearly identical to teacher assessment. Policy should support a shift toward more decentralized, learner-centered assessment frameworks. Educational policies often hesitate to formalize self-assessment due to concerns over grading inflation or unreliability; however, these data suggest that well-structured rubric development can justify integrating self-assessment into formal, high-stakes, or continuous evaluation policies. In terms of curriculum designs, this underscores the necessity of embedding explicit rubric training and formative self-evaluation loops directly into the syllabus layout. Rather than viewing assessment as a final task performed exclusively by the teacher, curricula should treat the mastery of assessment criteria as a core learning outcome. By doing so, institutions can foster sustainable assessment practices that alleviate teacher grading burdens while simultaneously cultivating long-term metacognitive autonomy in L2 writers.

4.4.5. 5.5 Limitations of the study

A key limitation of this study concerns the relatively small number of teacher raters ($n = 3$), which restricts the robustness of inferences regarding teacher assessment (TA) as a distinct facet within the Many-Facet Rasch Measurement (MFRM) framework. With such a limited number of raters, the estimation of teacher severity is more prone to individual rater effects, and the stability of rater parameter estimates may be reduced. This also has implications for the interpretation of rater separation reliability, as a small rater sample can constrain the variability needed for meaningful separation among rater groups. Consequently, the data showing teachers and students behave similarly in rating severity should be interpreted with caution, as it may partially reflect the limited dispersion of teacher severity estimates rather than a fully representative model of teacher rater variability. Apart from that, the use of PTW rubrics plays a crucial role in this study; however, the rubrics are designed for specific writing tasks, particularly short and structured formats such as emails and informal letters aligned with CEFR A2 level descriptors. Because of this, a single standardized essay prompt was used so that all students responded to the same task. This helped to ensure that differences in scores were mainly due to how raters applied the rubric rather than differences in the writing tasks themselves, making the results more comparable and easier to interpret.

5. Conclusion

The results showed no significant differences between teacher and student raters, suggesting that both groups applied the PTW rubrics in a consistent manner. In practical terms, student self-assessment was closely aligned with teacher ratings in the L2 writing context. More importantly, students were able to judge their own writing in ways that resemble trained raters, which suggests that self-assessment can play a meaningful role in helping learners understand writing criteria and gradually improve their writing skills. This also supports the idea that self-assessment is not only an evaluation tool but can be part of the learning process, especially in encouraging learner autonomy. From a measurement point of view, the findings suggest that clear rubrics together with rater training can help reduce

unnecessary variation in scoring. In the MFRM framework, this means that differences caused by rater severity were separated from differences in students' writing ability, allowing the scores to better reflect actual writing performance. Although some level of inconsistency is expected in human scoring, the variation observed in this study remained within acceptable limits, which supports the overall reliability of the assessment approach. Due to this study's limitations, future research should include a larger and more diverse pool of teacher raters to enable more stable estimation of rater severity and stronger generalisability of findings across instructional contexts, and also examine the applicability of the PTW rubrics across a wider range of writing genres and proficiency levels. Using multiple writing prompts would determine the stability of rater agreement across varying task demands. It will give more insight into the use of the primary trait type of rubric in diverse L2 writing assessment contexts.

References

- Ahmadi Shirazi, M. (2019). For a Greater Good: Bias Analysis in Writing Assessment. *Sage Open*, 9(1). <https://doi.org/10.1177/2158244018822377>
- Algeciras, K. A. M., & Oriden, P. E. (2023). Rubrics for assessment: Improving the role of the rater. *EDULEARN Proceedings*, 1, 8454–8460. <https://doi.org/10.21125/edulearn.2023.2215>
- Alghizzi, T. M., & Alshahrani, T. M. (2024). Effects of grading rubrics on EFL learners' writing in an EMI setting. *Heliyon*, 10(18), e36394. <https://doi.org/10.1016/j.heliyon.2024.e36394>
- Al-Nafjan, E., Alhawsawi, S., & Nasser Abu Dujayn, A. (2025). Exploring self-assessment in EFL academic writing: a study of undergraduate Saudi students utilizing a rubric. *Language Testing in Asia*, 15(1), 62. <https://doi.org/10.1186/s40468-025-00397-2>
- Ariakpomu, N. C., Holmes, M. J., & Ho, P. (2025). Measuring overall difference from a combination of attribute ratings with the many-facet Rasch model. *Food Quality and Preference*, 127, 105442. <https://doi.org/10.1016/j.foodqual.2025.105442>
- Asli, N. F., Mohd Matore, M. E. E., & Md Yunus, M. (2024). Construct validity of primary trait writing rubrics based on assessment use argument (AUA) validation framework. *Heliyon*, 10(22), e40053. <https://doi.org/10.1016/j.heliyon.2024.e40053>
- Azizah, N., Suseno, M., & Hayat, B. (2020). Severity-Leniency in Writing Assessment and Its Causal Factors. *KnE Social Sciences*. <https://doi.org/10.18502/kss.v4i14.7870>
- Baidoo-Anu, D., Asamoah, D., & Adusei, A. (2023). Teachers' beliefs and attitudes towards students' self assessment: A latent profile analysis. *International Journal of Educational Research Open*, 5, 100275. <https://doi.org/10.1016/j.ijedro.2023.100275>
- Cai, H., & Yan, X. (2024). Triangulating natural language processing (NLP)-based analysis of rater comments and many-facet Rasch measurement (MFRM): An innovative approach to investigating raters' application of rating scales in writing assessment. *Language Testing*, 41(2), 384–411. <https://doi.org/10.1177/02655322231210231>
- Chang, C.-C., & Tseng, J.-S. (2025). Reliability, validity, and strictness of adult learner online self-rating based on teacher and achievement criteria for different individual characteristics and academic achievement. *Studies in Educational Evaluation*, 84, 101431. <https://doi.org/10.1016/j.stueduc.2024.101431>
- El-Tallawy, S. N., Pergolizzi, J. V., Vasiliu-Feltes, I., Ahmed, R. S., LeQuang, J. K., El-Tallawy, H. N., Varrassi, G., & Nagiub, M. S. (2024). Incorporation of "Artificial Intelligence" for Objective Pain Assessment: A Comprehensive Review. *Pain and Therapy*, 13(3), 293–317. <https://doi.org/10.1007/s40122-024-00584-8>
- Engelhard, G., & Wind, S. A. (2019a). Introduction to the Special Issue on Rater-Mediated Assessments. *Journal of Educational Measurement*, 56(3), 475–477. <https://doi.org/10.1111/jedm.12221>
- Engelhard, G., & Wind, S. A. (2019b). Introduction to the Special Issue on Rater-Mediated Assessments. *Journal of Educational Measurement*, 56(3), 475–477. <https://doi.org/10.1111/jedm.12221>
- Esfandiari, R., & Myford, C. M. (2020). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, 18(2), 111–131. <https://doi.org/10.1016/j.asw.2012.12.002>

- Gonsalves, C., & Lin, Z. (2025). Clear in advance to whom? Exploring ‘transparency’ of assessment practices in UK higher education institution assessment policy. *Studies in Higher Education*, 50(7), 1454–1470. <https://doi.org/10.1080/03075079.2024.2381124>
- Han, C., Jiang, M., & Chen, Q. (2026). Rubricizing the assessment practice: A systematic review and meta-analysis of rubrics in rater-mediated assessment of language interpreting. *Language Testing*, 43(2), 197–232. <https://doi.org/10.1177/02655322251391233>
- Hasan, A., & Jones, B. (2024). Assessing the assessors: investigating the process of marking essays. *Frontiers in Oral Health*, 5. <https://doi.org/10.3389/froh.2024.1272692>
- Hong, W. C. H. (2025). Revisiting the impact of errors in L2/FL writing: A methodological review of five decades of research. *Research Methods in Applied Linguistics*, 4(3), 100269. <https://doi.org/10.1016/j.rmal.2025.100269>
- Huang, J., & Chen, G. (2022). Individualized feedback to raters in language assessment: Impacts on rater effects. *Assessing Writing*, 52, 100623. <https://doi.org/10.1016/j.asw.2022.100623>
- Humphry, S. M., & Heldsinger, S. (2019). A Two-Stage Method for Classroom Assessments of Essay Writing. *Journal of Educational Measurement*, 56(3), 505–520. <https://doi.org/10.1111/jedm.12223>
- Jin, K.-Y., & Eckes, T. (2026). Does Rating Time Predict Illusory Halo? A Mixture Rasch Facets Analysis of Halo Effects in Onscreen Assessments. *Journal of Educational and Behavioral Statistics*. <https://doi.org/10.3102/10769986261441662>
- Khamboonruang, A. (2023a). EFL Thai students’ rating performances in self- and peer-assessments of writing: A many-facets Rasch analysis. *LEARN Journal: Language Education and Acquisition Research Network*, 16(1), 221–245. <https://so04.tci-thaijo.org/index.php/LEARN/article/view/263440>
- Khamboonruang, A. (2023b). Detecting Differential Rater Severity in a High-Stakes EFL Classroom Writing Assessment: A Many-Facets Rasch Measurement Approach. *PASAA*, 66(1), 5–36. <https://doi.org/10.58837/CHULA.PASAA.66.1.1>
- Khojasteh, L., Kafipour, R., Pakdel, F., & Mukundan, J. (2025). Empowering medical students with AI writing co-pilots: design and validation of AI self-assessment toolkit. *BMC Medical Education*, 25(1), 159. <https://doi.org/10.1186/s12909-025-06753-3>
- Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, 38(4), 602–626. <https://doi.org/10.1177/0265532221994052>
- Lertsakulbunlue, S., & Kantiwong, A. (2024). Development of peer assessment rubrics in simulation-based learning for advanced cardiac life support skills among medical students. *Advances in Simulation*, 9(1), 25. <https://doi.org/10.1186/s41077-024-00301-7>
- Lestari, S. B., & Brunfaut, T. (2023). Operationalizing the reading-into-writing construct in analytic rating scales: Effects of different approaches on rating. *Language Testing*, 40(3), 684–722. <https://doi.org/10.1177/02655322231155561>
- Li, J., & Wang, Q. (2021). Development and validation of a rating scale for summarization as an integrated task. *Asian-Pacific Journal of Second and Foreign Language Education*, 6(1), 11. <https://doi.org/10.1186/s40862-021-00113-6>
- 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>
- Li, Y., Liao, X., & Jia, J. (2025). Effects of raters’ professional backgrounds on assessing interpreting quality: An exploratory mixed-methods investigation into rater behavior. *System*, 133, 103772. <https://doi.org/10.1016/j.system.2025.103772>
- Mohd Noh, M. F., & Mohd Matore, M. E. E. (2020). Rating Performance among Raters of Different Experience Through Multi-Facet Rasch Measurement (MFRM) Model. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 11(2), 147–162. <https://doi.org/10.21031/epod.662964>

- Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.941084>
- Neittaanmäki, R., & Lamprianou, I. (2024). All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency. *Language Testing*, 41(3), 606–626. <https://doi.org/10.1177/02655322241239362>
- Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, 30, 100329. <https://doi.org/10.1016/j.edurev.2020.100329>
- Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-efficacy: a Meta-analytic Review. *Educational Psychology Review*, 35(4), 113. <https://doi.org/10.1007/s10648-023-09823-4>
- Roberts, S. K., & Whitfield, E. (2025). Exploring Current State Writing Evaluation Rubrics for Early Grades. *Early Childhood Education Journal*, 53(6), 1951–1960. <https://doi.org/10.1007/s10643-024-01823-5>
- Royal, K. D. (2014). Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences. *Psychometrika*, 79(4), 733–735. <https://doi.org/10.1007/s11336-013-9398-1>
- Safari, F., & Ahmadi, A. (2025). L2 learners' challenges in integrated writing tasks: implications for writing teachers and developers of diagnostic assessments. *The Language Learning Journal*, 53(1), 53–70. <https://doi.org/10.1080/09571736.2023.2301604>
- Şahan, Ö., & Razi, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*, 37(3), 311–332. <https://doi.org/10.1177/0265532219900228>
- Stranford, S. A. (2024). Fostering student agency and motivation: co-creation of a rubric for self-evaluation in an ungraded course. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1213444>
- Su, Y., & Shin, S.-Y. (2024). Comparing two formats of data-driven rating scales for classroom assessment of pragmatic performance with roleplays. *Language Testing*, 41(2), 357–383. <https://doi.org/10.1177/02655322231210217>
- Tanaka, M. (2025). Friendship bias in peer assessment of EFL oral presentations. *Assessment in Education: Principles, Policy & Practice*, 32(5–6), 515–537. <https://doi.org/10.1080/0969594X.2025.2570248>
- Tian, W., & Tong, Y. (2026). A Literature Review of Effects of Rubric on Student Self-Assessment and Peer Assessment in Higher Education. *Journal of Language, Culture and Education*, 3(2), 1–13. <https://doi.org/10.70267/jlce.2026.v3n2.0113>
- Uto, M., Tsuruta, J., Araki, K., & Ueno, M. (2024). Item response theory model highlighting rating scale of a rubric and rater–rubric interaction in objective structured clinical examination. *PLOS ONE*, 19(9), e0309887. <https://doi.org/10.1371/journal.pone.0309887>
- Vasileiadou, D., & Karadimitriou, K. (2021). Examining the impact of self-assessment with the use of rubrics on primary school students' performance. *International Journal of Educational Research Open*, 2, 100031. <https://doi.org/10.1016/j.ijedro.2021.100031>
- Veloo, A., Hashima Abd Aziz, N., & Yaacob, A. (2018). The Most Suitable Scoring Method to Assess Essay Writing in ESL Classrooms. *Advances in Language and Literary Studies*, 9(4), 19. <https://doi.org/10.7575/aiac.all.v.9n.4p.19>
- White, M., Göllner, R., & Kleickmann, T. (2025). Reconsidering the measurement of teaching quality using a lens model: measurement dilemmas and trade-offs. *School Effectiveness and School Improvement*, 36(2), 192–210. <https://doi.org/10.1080/09243453.2025.2482583>
- Wu, M. Y., Steinkrauss, R., & Lowie, W. (2023). The reliability of single task assessment in longitudinal L2 writing research. *Journal of Second Language Writing*, 59, 100950. <https://doi.org/10.1016/j.jslw.2022.100950>
- Yan, Z. (2022). *Student Self-Assessment as a Process for Learning*. Routledge. <https://doi.org/10.4324/9781003162605>
- Yu, S., Zhang, Y., Liu, C., & Lee, I. (2025). From theory to practice: Understanding the long-term impact of an L2 writing education course on writing teachers. *Language Teaching Research*, 29(6), 2549–2582. <https://doi.org/10.1177/13621688221130852>

- Zheng, B., He, Q., & Lei, J. (2024). Informing factors and outcomes of self-assessment practices in medical education: a systematic review. *Annals of Medicine*, 56(1). <https://doi.org/10.1080/07853890.2024.2421441>
- Zheng, Y., Liu, C., & Yu, S. (2026). “How should I assess their writing? It’s a headache for me.”: Understanding teacher assessment literacy in the collaborative writing context. *Teaching and Teacher Education*, 171, 105343. <https://doi.org/10.1016/j.tate.2025.105343>