Raters' Rating Quality in Assessing Students' Assignment: An Application of Multi-Facet Rasch Measurement

Chan Kuan Loong, Abdul Jalil Mohamad, Mohd Syafiq bin Zainuddin, Navindran a/l Ramanujan, Ikmal Hisham bin Mat Idera

Jabatan Ilmu Pendidikan, Institut Pendidikan Guru Kampus Tengku Ampuan Afzan

DOI: https://dx.doi.org/10.47772/IJRISS.2025.910000193

Received: 07 October 2025; Accepted: 14 October 2025; Published: 07 November 2025

ABSTRACT

Assessing students' assignments is essential as it reflects students' understanding and achievements. This study evaluates the marking quality among lecturers at the Teacher Education Institute (IPG) using the Multi-Facet Rasch Measurement (MFRM) model. Two hundred thirty-two students from the Postgraduate Diploma in Education (PDPP) program submitted written assignments, which were assessed by experienced lecturers. The analysis was conducted using FACETS 4.1.1 software, involving three main facets: candidates, raters, and assessment criteria. The findings indicate that the instrument used is valid in terms of construct and meets unidimensionality requirements. The reliability value of the raters was high (0.81), and the rater separation index exceeded the set threshold (2.09), indicating stability in marks given by lecturers. However, two raters were identified as showing misfit and overfit patterns respectively, suggesting inaccuracies in scoring. The Wright map and unexpected response analysis also revealed differences in the severity among raters and potential bias. These findings are valuable for the IPG in improving monitoring of inter-rater reliability and marking consistency. This study also shows that MFRM can provide comprehensive information and contribute to understanding the analysis of assessor consistency with quantitative evidence. MFRM is a suitable alternative model to overcome the limitations of the Classical Test Theory (CTT) statistical model, especially in analyses involving multiple raters.

Keywords: Rating quality, lecturers, MFRM

INTRODUCTION

During the assessment process, scoring is a fundamental component that directly reflects students' cognitive performance, learning progression, and academic achievement. Particularly in open-ended tasks, the quality of scoring plays a pivotal role in shaping not only the validity and reliability of the results but also students' motivation and engagement. As such, the scoring process must be guided by precise criteria, applied with fairness, and executed with consistency to ensure trustworthy inferences about student learning [1], [2].

Open-ended responses, is an effective way in capturing students' higher-order thinking, present challenges due to their subjective and multifaceted nature. Unlike multiple choice items, these tasks require evaluators to interpret diverse modes of expression, thereby introducing ambiguity and potential bias. To address this, educators employ various scoring approaches—holistic, analytic, and multi-trait—each offering distinct advantages depending on the task's learning goals [3], [4]. However, the effectiveness of these methods hinges on the clarity of rubrics and the scorers' judgment capacity.

Despite the availability of structured rubrics, scoring practices remain susceptible to rater-related distortions, including severity, leniency, central tendency, and the halo effect. These effects are especially pronounced in open-ended assessments, where subjective interpretation can overshadow objective evaluation [5], [6]. Moreover, extraneous variables such as students' handwriting, language style, or perceived effort may further contaminate scoring validity, raising critical concerns about fairness and accuracy [7].

In Malaysian Teacher Education Institutes (IPG), moderation processes are employed to standardise assessment judgments across lecturers. While moderation can reduce overt discrepancies, it lacks the psychometric precision necessary to diagnose latent scoring inconsistencies. Furthermore, current moderation practices often do not incorporate statistical modelling or systematic training that could enhance inter-rater reliability and provide diagnostic feedback [8], [9].

Although psychometric models such as Rasch analysis have demonstrated utility in evaluating rater behaviour and rubric functioning [10], [11], their application within the IPG context remains limited. The literature lacks empirical investigations that systematically examine marking quality, rater bias, and score reliability among IPG lecturers. Therefore, this study seeks to fill this gap by exploring the nature and extent of rater effects in the assessment of open-ended student work, while also evaluating the potential of Rasch-based psychometric approaches to enhance assessment fidelity and scoring consistency in teacher education.

Research Objective

Assess the validity and reliability of each facet.

Determine the level of severity of raters in evaluating student assignments.

Evaluate the impact of examiner bias in measuring student performance using the generated rubric.

LITERATURE REVIEW

Assessment is an integral component of educational systems, as it provides insights into students' achievement levels and offers feedback to teachers regarding the effectiveness of their instructional methods. High-quality assessment serves as the foundation for a fair and valid evaluation system. Nevertheless, issues of subjectivity and inconsistency among assessors continue to pose significant challenges, particularly within teacher education institutions (Institut Pendidikan Guru, IPG), which emphasise performance-based and authentic assessment. In this context, the validity and reliability of marking heavily depend on transparency, consistency, and fairness in the grading process [12], [13].

Assessment validity refers to the extent to which an assessment accurately measures the intended learning outcomes [14]. In the context of IPGs, this validity can be bolstered by implementing explicit and comprehensive scoring rubrics, alongside thorough training for raters. Nevertheless, empirical studies have indicated that even with the application of rubrics, discrepancies in scoring among raters continue to exist, especially in subjective evaluations like written tasks and oral presentations [15]. Such discrepancies predominantly stem from variations in raters' experiences, linguistic proficiencies, and their understanding of the rubrics utilised.

One critical issue closely linked to assessment reliability is rater bias, which occurs when raters exhibit scoring patterns that deviate from established assessment criteria. This phenomenon is known as differential rater functioning (DRF), where raters display varying degrees of strictness or leniency depending on the task type, the student being assessed, or the assessment domain. Research by [16] demonstrates that rater bias can significantly affect students' actual scores, thus compromising the fairness of assessments. Factors such as experience and training also play significant roles. For instance, Sureeyatanapas et al. [17] found that inexperienced raters displayed greater variability in strictness prior to training but became more consistent following training interventions.

Furtehrmore, evidence indicates that raters tend to score domains perceived as difficult more leniently, while rating domains they consider easy more strictly [18], [19]. This observation implies that evaluators' interpretations of domain difficulty may shape their scoring behaviors, thereby influencing the transparency of assessments. Additionally, evaluators might fall prey to the "halo effect," where overall ratings are subject to inflation or deflation based on preliminary general perceptions of student performance [20]. In this context, superficial readings of rubrics devoid of comprehensive understanding have been recognized as a fundamental factor contributing to scoring variability [21].

Modern measurement approaches such as the Many-Facet Rasch Measurement (MFRM) model offer a systematic and transparent framework to address these issues. MFRM allows for the analysis of interactions between various facets of the assessment system, such as candidates, items, raters, and scoring criteria [22]. Unlike the basic Rasch model, which analyses only two facets (candidates and items), MFRM is more suitable for polytomous scores and multi-raters' assessments. This model can identify rater severity or leniency and adjust student scores to ensure fairness and reflect their actual abilities.

Through MFRM, the parameters of each facet are calibrated based on logits, enabling the analysis of rater consistency, item difficulty, and the effectiveness of score categories in the rubric. The model also provides statistics such as the separation reliability index and fit statistics to assess how well the data align with the model. Additionally, MFRM generates fair scores, which are adjusted student scores that account for the influence of rater bias, making the assessment more objective and credible [23], [24].

Previous studies demonstrated the effectiveness of MFRM across various evaluative frameworks. Elder et al. [32] showed that instructional interventions can enhance the precision and uniformity of scoring. In a parallel vein, the research conducted by Khabbazbashi et al. [33] concerning the CEFR assessment revealed that the fairness of oral evaluations was augmented when raters received training informed by MFRM data. Furthermore, within the realm of raters training, MFRM analysis is a tool for evaluating the effectiveness of assessment rubrics, pinpointing additional training requirements, and refining overall grading methodologies.

In summary, findings from multiple studies emphasise the need for a systematic, fair, and transparent scoring system within the context of teacher education institutions (IPG) and higher education. Variations in rater strictness, unconscious biases, and challenges in interpreting rubrics have significant implications for the validity and reliability of assessments. Therefore, using approaches like MFRM enhances the integrity of the assessment system and contributes to developing professionalism among educators and raters.

RESEARCH METHODOLOGY

Research Design

Before conducting the actual study, a mapping of assessment elements was carried out to facilitate the Multi-Facet Rasch Measurement (MFRM) analysis using FACET Minifac 64 software. In mapping the data for multi-rater assessments, special attention must be given to the connectedness among the assessed facets to ensure a comprehensive analysis of each facet element. Engelhard Jr et al. [13] noted several assessment designs suitable for MFRM analysis, including a fully-crossed rating design, a linked rating design, and a disconnected rating design.

In this study, the researcher selected a linked rating design. The linked design allows raters to assess only a subset of candidates and items while maintaining data connectedness through a systematic network structure. This method conserves resources and enhances assessment reliability through cross-rater comparisons. In this study, three scripts were randomly selected for anchoring purpose. All lecturers marked their scripts to establish connectedness within the assessment system. This approach ensures accurate and comprehensive analysis without requiring all raters to evaluate every script. Table 1 shows how student scores were collected from raters to ensure the existence of connectedness between each facet of the assessment.

Table 1: Assessment Mapping

Script	**Rater**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Anchor 1	x	x	x	x	x	x	x	x	x	x
Anchor 2	x	x	x	x	x	x	x	x	x	x

Anchor 3	x	x	x	x	x	x	x	x	x	x
1-229	Each script was rated by one lecturer									

Participant

This study involved 232 students who enrolled in the first semester of the Postgraduate Diploma in Education Program (PDPP). Of this number, 65 participants were male and 167 participants were female. The participants come from diverse academic backgrounds; however, they all share the goal of enhancing their understanding of pedagogical concepts and their application in teaching and learning practices

Rater

This study involved 10 lecturers from IPGM Campus Tengku Ampuan Afzan (IPGM KTAA) as raters (three male and seven female). All raters are lecturers with at least ten years of teaching experience. However, only 3 of them have specific experience teaching this subject. To ensure consistency in assessment, all raters were provided with a clear scoring rubric and training to ensure that the evaluation process was consistent and fair. The diversity in the raters' backgrounds provided a broader perspective in the assessment process, which contributed to the effectiveness of this study in evaluating the understanding and application of pedagogical concepts among the study participants.

Instrument

232 PDPP students were assigned to produce an academic writing of 1600 words within 27 days. This task focused on how to design a meaningful learning plan. According to the established scoring rubric, the written work will be assessed based on Cognitive Skills (30%) and Personal Skills (10%).

Script Collection and Marking

After 27 days, all scripts were collected from the students for the assessment process. Before the marking process began, all raters attended a marking training session to ensure a consistent understanding of the assessment criteria and to enhance the uniformity in scoring. Following this, three scripts were randomly selected as anchor scripts and distributed to all raters for joint evaluation. Once this collaborative marking process was completed, each rater evaluated the student scripts from their respective class based on the provided scoring rubric.

Data Analysis

This study employed Multi-Facet Rasch Measurement (MFRM) to assess the quality of lecturer marking. This method was used to analyse regression patterns among raters who evaluated the students' writings, ensuring fairness and reliability in the scoring process. Four main elements were analysed in this study using FACETS Minifac Version 4.1.1 software. First, an unidimensionality test was applied to assess the construct validity and the evaluated items. Second, the rater validity was analysed using Outfit Mean Square (MNSQ) in the Rater Measurement Report to ensure consistency in the marking process. Third, a Wright Map analysis was conducted to determine each rater's severity level or leniency in scoring. Finally, the Outlier Response element was used to measure the potential bias in marking related to the quality of student writing.

RESEARCH FINDINGS

Unidimensionality

Before rater severity can be analysed comprehensively, the items used must be validated to ensure they accurately measure the intended construct.

Diagram 1: Unidimensionality

		Count	Mean	S.D.
Responses non-extreme estimable	=	504	3.80	0.65
Responses in one extreme score	=	14	5.00	0.00
All Responses	=	518	3.83	0.67
Count of measurable responses	=	518		
Raw-score variance of observations	=	0.421	100.00%	
Variance explained by Rasch measures	=	0.245	58.27%	
Variance of residuals	=	0.176	41.73%	

The items in the student assignment assessment showed a variance explained by measure of 58.27%. This value exceeds 40%, indicating that both domains meet the unidimensionality requirement and fall into the category of good quality [27]. Furthermore, this value approaching 60% suggests an excellent level of unidimensionality. Additionally, the obtained eigenvalue is 0.245, well below the threshold value of 3. This indicates that there are no problematic items within the assessment instrument [27], [28]. Both data points confirm that the instrument used is valid and meets the fundamental requirements of MFRM, thus enabling more accurate and data-driven analysis of rater severity.

Rater Validity

Rater validity was conducted to determine the accuracy and efficiency of raters in evaluating student assignments [29]. This statistic also provides information about the level of consistency among raters in ranking students according to their abilities [5]. It assesses how well the ratings produced by the raters align with the expectations of the measurement model [30] by analysing any gaps between observed scores and expected scores [31]. The Infit MNSQ index plays a role in identifying the fit of the data to the model, particularly in detecting outliers in the assessment [32]. Generally, MNSQ values within an acceptable range are between 0.50 and 1.50 [33].

Figure 2 shows the statistical measures for the rater facet, including the level of severity (measure), standard error (SE) and the Infit and Outfit Mean Square (MNSQ) values. The level of rater severity refers to their tendency to assign scores either more leniently or more strictly towards students [34]. In this study, the level of rater severity ranged from -2.28 logits (SE = 0.28) for Rater 5, the most lenient, to 0.51 logits (SE = 0.34) for Rater 7, the most severe. The standard error (SE) values indicate the precision of each estimate of severity, where smaller SE values reflect higher measurement accuracy [33].

According to Eckes [45], there are two types of fit statistics: misfit and overfit. A fit statistic below 0.5 is considered overfit, which indicates that the raters do not exhibit enough variation in their assessments. Conversely, a fit statistic above 1.5 indicates misfit, or excessively high inconsistency [32]. Based on Figure 2, it was found that Rater 8 was overfit (0.14). This suggests that this rater did not show enough variation in their ratings. Rater 5, on the other hand, showed a misfit value (1.86). According to Eckes [46], misfit among raters is more concerning than overfit because high-scoring inconsistency can undermine the assessment results' reliability and validity.

Table 1 shows the scores produced by Rater 5 and Rater 8, who failed to mark student assignments consistently. For example, the average score difference produced by Rater 5 is inconsistent. There was a slight score difference for Student 206, only 0.32 (4.50-4.18), but a significant difference for Student 194, with a discrepancy of 0.86 (3.50-1.94). Rater 8, meanwhile, was observed to mostly use ratings of 3 and 4, with no extreme scale values applied.

The findings of this study align with research by [2], who used MFRM to assess the marking quality of 164 teachers in evaluating English language oral exams. Of the 164 teachers, two were found to be misfit and two overfit. This study also showed that misfit teachers tended to produce inconsistent marking patterns, as they were at times stringent and at other times lenient. Overfit teachers exhibited a central tendency effect, using mostly middle-range scale categories with no extreme scale values used in their ratings.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Rater
140	38	3.68	3.61	.51	.34	1.23	.8	1.88	2.1	.43	.78	.81	44.4	63.0	7	7
277	70	3.96	3.69	.17	.30	.51	-2.6	.31	-3.0	1.36	.79	.74	63.0	63.8	6	6
337	84	4.01	3.72	.01	.33	.14	-4.9	.07	-5.2	1.39	.94	.74	63.0	63.9	8	8
174	46	3.78	3.78	-.30	.31	1.03	.1	.99	.0	.90	.77	.80	63.0	63.7	9	9
188	50	3.76	3.78	-.30	.33	.55	-1.8	.54	-1.4	1.30	.86	.80	50.0	63.7	10	10
142	38	3.74	3.78	-.30	.32	.61	-2.0	.61	-1.8	1.55	.80	.78	59.3	63.7	3	3
229	56	4.09	3.80	-.42	.33	.63	-1.9	.61	-1.2	1.32	.84	.77	57.4	63.5	4	4
139	42	3.31	3.90	-.94	.29	1.32	1.2	2.20	3.1	.27	.51	.77	59.3	61.6	2	2
158	40	3.95	3.92	-1.10	.38	.82	-.5	.96	.0	1.04	.82	.81	44.4	60.7	1	1
202	54	3.74	4.11	-2.28	.28	1.86	3.8	3.31	6.1	-.75	.76	.73	37.0	50.7	5	5
198.6	51.8	3.80	3.81	-.49	.32	.87	-.8	1.15	-.1		.79					Mean (Count: 10)
62.3	14.3	.21	.13	.75	.03	.47	2.3	.96	3.1		.11					S.D. (Population)
65.7	15.1	.22	.14	.79	.03	.50	2.5	1.01	3.3		.11					S.D. (Sample)

Model, Populn: RMSE .32 Adj (True) S.D. .67 Separation 2.09 Strata 3.11 Reliability (not inter-rater) .81
Model, Sample: RMSE .32 Adj (True) S.D. .72 Separation 2.22 Strata 3.30 Reliability (not inter-rater) .83
Model, Fixed (all same) chi-squared: 61.7 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-squared: 7.9 d.f.: 8 significance (probability): .44
Inter-Rater agreement opportunities: 270 Exact agreements: 146 = 54.1% Expected: 166.9 = 61.8%

Figure 2: Rater Measurement Report (Fit Statistic)

Table 1: Example of Scores Produced by Misfit Rater (Rater 5) Exhibiting Contamination Effect and Central Tendency Effect (Rater 8)

Rater	Student	Item		Average marks	Fair-Average		Rater	Student	Item		Avrage mark	Fair-Average
		HP2	HP9						HP2	HP9		
5	3	4	4	4.00	3.20		**8**	3	3	3	3.00	3.20
	12	4	4	4.00	4.28			12	4	4	4.00	4.28
	22	5	5	5.00	3.97			22	4	4	4.00	3.97
	193	3	3	3.00	2.17			78	4	4	4.00	4.04
	194	3	4	3.50	2.64			79	4	4	4.00	4.04
	195	3	4	3.50	2.64			80	4	4	4.00	4.04
	196	3	4	3.50	2.64			81	4	4	4.00	4.04
	197	3	4	3.50	2.64			82	4	4	4.00	4.04
	198	3	5	4.00	3.66			83	4	4	4.00	4.04
	199	3	4	3.50	2.64			84	4	5	4.50	4.58
	200	3	4	3.50	2.64			85	4	4	4.00	4.04
	201	3	4	3.50	2.64			86	4	5	4.50	4.58
	202	3	4	3.50	2.64			87	4	4	4.00	4.04
	203	3	4	3.50	2.64			88	4	4	4.00	4.04
	204	3	4	3.50	2.64			89	4	4	4.00	4.04
	205	3	4	3.50	2.64			90	4	4	4.00	4.04

	206	4	5	4.50	4.18			91	4	4	4.00	4.04
	207	3	3	3.00	2.17			92	4	4	4.00	4.04
	208	3	4	3.50	2.64			93	4	4	4.00	4.04
	209	3	5	4.00	3.66			94	4	4	4.00	4.04
	210	3	5	4.00	3.66			95	4	4	4.00	4.04
	211	3	5	4.00	3.66			96	4	4	4.00	4.04
	212	3	5	4.00	3.66			97	4	4	4.00	4.04
	213	3	4	3.50	2.64			43	4	4	4.00	4.04
	214	4	5	4.50	4.18			44	4	4	4.00	4.04
	215	3	4	3.50	2.64			45	4	5	4.50	4.58
	216	3	5	4.00	3.66			46	4	4	4.00	4.04
								47	4	4	4.00	4.04
								48	4	4	4.00	4.04
								49	4	4	4.00	4.04
								50	4	4	4.00	4.04
								51	4	4	4.00	4.04
								52	4	4	4.00	4.04
								53	4	4	4.00	4.04
								53	4	4	4.00	4.04
								54	4	4	4.00	4.04
								55	4	4	4.00	4.04
								56	4	4	4.00	4.04
								57	4	4	4.00	4.04
								58	4	4	4.00	4.04
								59	4	4	4.00	4.04

Item Validity

Based on the statistical analysis, both domains in this study meet the requirements of the Rasch model and show a high degree of fit. The Infit MNSQ values for both domains fall within the acceptable range (0.71–1.22), while the Outfit MNSQ values range from 0.79 to 1.19. Additionally, the Outfit Zstd values for both domains are within the ±2.0 range. In terms of criterion difficulty, HP2 is the more difficult criterion (1.49 logits, SE = 0.16),

compared to HP9 (-1.49 logits, SE = 0.13). The measurement accuracy can be determined based on the standard error (SE) recorded in Figure 3. The standard error for both criteria falls within the range of 0.13 to 0.16, indicating good precision. Therefore, it can be concluded that all assessed domains are appropriately valued and align with the Rasch measurement model, thus supporting the instrument's validity.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Aitem
1002	259	3.87	3.91	1.49	.16	1.22	2.3	1.34	2.0	.79	.67	.72	1 HP2
984	259	3.80	3.90	-1.49	.13	.71	-2.9	.74	-2.0	1.19	.80	.75	2 HP9
993.0	259.0	3.83	3.91	.00	.15	.96	-.3	1.04	.0		.74		Mean (Count: 2)
9.0	.0	.03	.01	1.49	.02	.26	2.6	.30	2.0		.07		S.D. (Population)
12.7	.0	.05	.01	2.11	.02	.36	3.7	.42	2.9		.09		S.D. (Sample)

Model, Populn: RMSE .15 Adj (True) S.D. 1.48 Separation 10.15 Strata 13.86 Reliability .99
Model, Sample: RMSE .15 Adj (True) S.D. 2.10 Separation 14.39 Strata 19.51 Reliability 1.00
Model, Fixed (all same) chi-squared: 207.9 d.f.: 1 significance (probability): .00

Figure 3: Item Measurement Report (Domain Statistic)

Reliability

The reliability of Rasch analysis can be evaluated using the separation index and the reliability index. The separation index indicates how well the elements within each facet can be differentiated so that each facet is clearly defined [37]. If the separation index exceeds 2, it is considered a good and acceptable value [38]. The Rasch reliability index ranges from 0 to 1. A value approaching 1 indicates that the model is good and effective with a high level of consistency [39].

Based on Figure 2, the reliability value for the raters is high, at 0.81. The rater separation index is also good, as it exceeds 2, at 2.09. The significant value of p = 0.00 indicates that there is a significant difference in rater severity. This shows that raters exhibit varying severity levels when assessing student assignments. The actual percentage of rater agreement is 54.1%, which is not far from the expected percentage of rater agreement (61.8%). This indicates that the assessments made by the experts are neither homogeneous nor perfect, but are still considered good because they align with the expectations of the Rasch Model.

Wright Map

Figure 4 shows the Wright map, which illustrates the positioning of each facet involved in the analysis. The first column represents the measurement scale in logits, followed by the second column indicating the distribution of student ability levels. The third column displays the difficulty levels of the assessment criteria, while the fourth reflects the severity levels of the raters. Finally, the fifth and sixth columns represent the rubric score scale. By aligning these three facets—students, items, and raters—on the same measurement scale (expressed in logit units), the quality of each facet can be analysed and compared on a standard metric.

Within the rater severity column, Rater 2 and Rater 1 were positioned closely together, indicating similar levels of severity. Likewise, Raters 4, 9, 3, 10, 8, 6, and 7 form a distinct cluster, suggesting their severity levels are also closely aligned. However, Rater 5 appears markedly lower than the others, indicating that this individual was the most lenient in awarding scores. In contrast, Rater 7 is positioned at the highest point on the Wright map, signifying that this rater exhibited the greatest severity among all raters.

Although differences in severity levels exist among the raters, the variation is not substantial, as eight of them fall within the logit range of -1.0 to 1.0. According to Eckes [39], raters with severity estimates ≥ 1.0 logits are classified as severe, while those with ≤ -1.0 logits are considered lenient. In this study, only two raters fall into the lenient category: Rater 1 (-1.10 logits) and Rater 5 (-2.28 logits).

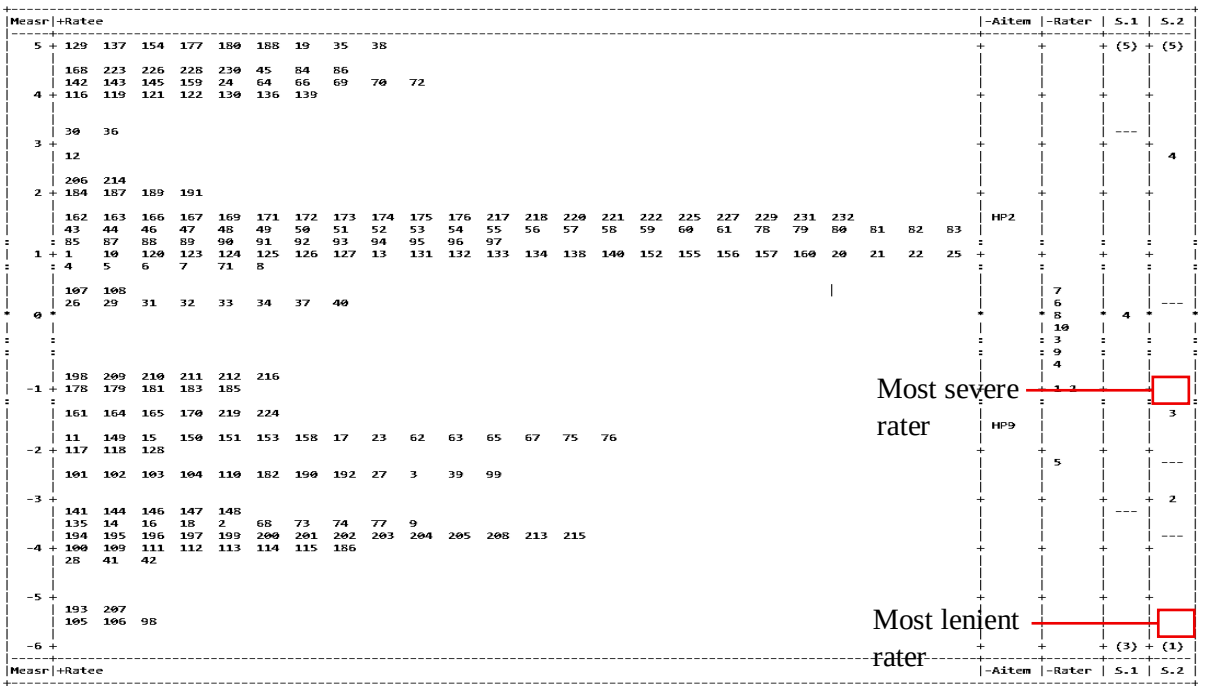

Figure 4: Distribution of Each Facet (Wright Map

Unexpected Responses

Unexpected responses in the Multi-Facet Rasch Model (MFRM) reveal discrepancies between observed and expected scores, helping to identify rater bias or inconsistency. Analyzing these responses provides insights into the fairness and adherence of raters to assessment criteria.

Based on Figure 5, there were 11 responses in which raters awarded lower-than-expected scores (under-value), and 16 responses with higher-than-expected scores (over-value). Rater 5, in particular, exhibited thirteen instances of unexpected or biased responses. This pattern suggests that Rater 5 encountered difficulties interpreting and consistently applying the assessment rubric. An intervention in the form of retraining should align this rater's perceptions and understanding of the construct being assessed, in accordance with the principles of valid and reliable assessment.

Cat	Score	Exp.	Resd	StRes	Nu	Ra	Num	Rat	N	Ait	Sequence
4	4	3.1	.9	4.0	2	2	98	98	1	HP2	47
4	4	3.1	.9	4.0	2	2	105	105	1	HP2	61
4	4	3.1	.9	4.0	2	2	106	106	1	HP2	63
5	5	4.0	1.0	3.6	4	4	22	22	1	HP2	125
3	3	4.0	-1.0	-3.6	5	5	198	198	1	HP2	193
3	3	4.0	-1.0	-3.6	5	5	209	209	1	HP2	215
3	3	4.0	-1.0	-3.6	5	5	210	210	1	HP2	217
3	3	4.0	-1.0	-3.6	5	5	211	211	1	HP2	219
3	3	4.0	-1.0	-3.6	5	5	212	212	1	HP2	221
3	3	4.0	-1.0	-3.6	5	5	216	216	1	HP2	229
3	3	4.0	-1.0	-3.6	7	7	191	191	1	HP2	335
5	5	4.0	1.0	3.3	1	1	22	22	1	HP2	5
3	3	4.1	-1.1	-3.1	1	1	22	22	2	HP9	6
5	5	4.1	.9	3.1	7	7	12	12	1	HP2	303
5	5	4.0	1.0	3.1	7	7	191	191	2	HP9	336
5	5	4.0	1.0	3.0	5	5	198	198	2	HP9	194
5	5	4.0	1.0	3.0	5	5	209	209	2	HP9	216
5	5	4.0	1.0	3.0	5	5	210	210	2	HP9	218
5	5	4.0	1.0	3.0	5	5	211	211	2	HP9	220
5	5	4.0	1.0	3.0	5	5	212	212	2	HP9	222
5	5	4.0	1.0	3.0	5	5	216	216	2	HP9	230
3	3	4.0	-1.0	-3.0	10	10	22	22	2	HP9	474
3	3	3.9	-.9	-2.7	7	7	22	22	1	HP2	305
3	3	3.9	-.9	-2.3	7	7	22	22	2	HP9	306
5	5	4.2	.8	2.2	9	9	12	12	1	HP2	425
5	5	4.2	.8	2.2	10	10	12	12	1	HP2	471
5	5	4.2	.8	2.1	5	5	22	22	1	HP2	181
Cat	Score	Exp.	Resd	StRes	Nu	Ra	Num	Rat	N	Ait	Sequence

Figure 5: Unexpected Responses

DISCUSSION

The findings of this study indicate that the data obtained are consistent with the Rasch Model. The constructed assignment questions demonstrate strong unidimensionality, as evidenced by the percentage of variance explained by the Rasch measures (58.26%). This suggests that the instrument functions effectively in measuring students' abilities. In terms of reliability, all relevant indices—including reliability index, strata, and separation—show excellent values, reflecting stable measurement within the context of a multi-rater assessment.

These results demonstrate that the Multi-Facet Rasch Model (MFRM) enables a more comprehensive and in-depth analysis of the validity and reliability of an assessment instrument or rubric, compared to traditional approaches such as Classical Test Theory (CTT) [5]. A study by Abd Rahman et al. [21] employed the MFRM to evaluate the validity and reliability of the Student Mathematical Process Rubric (ProM3), which was developed based on 29 criteria across five dimensions. The assessment involved three key facets: student ability, rater severity, and item difficulty. Their findings confirmed that the MFRM could accurately evaluate the function of each facet within the assessment system. Furthermore, MFRM analysis provided comprehensive diagnostic information to support the refinement of rubric design, particularly in assessing students' mathematical processes. Thus, this approach enhances rubric quality and strengthens fairness and transparency in criterion-referenced assessment.

One of the distinctive strengths of the MFRM lies in its capacity to generate person-centred statistics, including those pertaining to raters, candidates, and assessment tasks. The MFRM successfully identified differences in rater severity and leniency in the present study. This can be showed through the rater facet of the Wright Map, illustrates differences in the logit measures of rater stringency. Overall, most raters were within a moderate range and did not fall into the category of being excessively severe or lenient. However, two raters were identified as displaying a tendency towards leniency in their scoring. These findings provide clear evidence of variation in rater behaviour, thereby supporting the accuracy and utility of the MFRM in assessing inter-rater consistency.

This finding is consistent with previous research by Erguvan et al. [44], who reported that while raters tended to apply rubrics consistently throughout the assessment of student assignments, they still differed in their levels of severity and leniency. Moreover, raters were inclined to assign scores of 70 and 80 more frequently than other score ranges, reflecting a possible central tendency effect. This phenomenon is often associated with raters' reluctance to provide justification for awarding extreme scores—particularly in the context of high-stakes assessment [41].

Another notable advantage of the MFRM is its ability to detect unexpected responses or biases among raters. Overall, the study found that most raters performed their evaluations professionally. Regarding bias, the model identified one rater—Rater 5—as exhibiting 13 biased interactions. Such bias further supports the finding that rater severity is associated with score inaccuracy. Raters who tend to score more leniently were found to be less precise in their judgements. Bias often arises when raters fail to give due attention to the established assessment criteria for each aspect being evaluated, impairing their ability to assign scores objectively [42]. Mohd Zabidi et al. [48] also identified subjective interpretation of rubric criteria as contributing to bias. To mitigate such biases, it is recommended that regular rater training be conducted to minimise overly subjective assessment practices.

CONCLUSION

The findings of this study reveal variability in lecturers' assessments, which may lead to dissatisfaction among students, particularly when more severe raters assess them. The presence of unexpected responses and evidence of bias further highlights the need for ongoing rater training to enhance consistency in assessment [40], [41]. These findings may also serve as valuable material in rater training programmes, fostering greater awareness among assessors regarding their levels of severity and promoting alignment in assessment standards.

In addition, multi-rater analysis using the Multi-Facet Rasch Model (MFRM) offers a comprehensive overview of rater consistency. MFRM can be utilised to identify and mitigate biased assessments, detect underperforming raters, and uncover sources of bias in the evaluation process. Unlike Classical Test Theory (CTT), which

primarily focuses on group-based statistics, MFRM provides more granular information regarding individual rater tendencies and severity levels. This, in turn, contributes to strengthening assessment validity [42].

Furthermore, the study found that including more than two criteria (specifically HP2 and HP5) is essential for achieving a broader distribution of item difficulty and measuring student ability more effectively. A limited number of criteria may result in an imbalanced distribution within a narrow ability range. Therefore, expanding the number of criteria is necessary to ensure a more even spread of item difficulty levels. This can be accomplished by deconstructing the existing criteria into more specific and distinct subcomponents.

In addition, these findings underscore the importance of ongoing rater calibration and rubric refinement in maintaining assessment validity. Visual representations such as Wright Maps and rater severity distributions provide transparent evidence of how raters function within the measurement framework. Systematic rater training—especially for criteria involving subjective judgement—can help reduce bias and improve shared interpretation of rubrics [5]. The integration of both quantitative and qualitative rater feedback offers a more holistic approach to improving rating quality and advancing fairness in performance-based assessment.

REFERENCES

1. M. H. Mazarul Hasan, Z. Norazimah, and M. Suhaila, “Readiness level of primary school teachers in Klang district, Selangor in the implementation of in-class assessment from the aspect of knowledge,” International Journal of Modern Education, vol. 3, no. 9, pp. 1–8, 2021.
2. M. E. @ E. Mohd Matore and M. F. Mohd Noh, Kualiti Penandaan Guru Dalam Pentaksiran Pendidikan. Penerbit UKM, 2023.
3. O. Ribut, Y. Pradana, A. Mashuri, L. S. Nirawati,) Stkip, and M. Ngawi, “Pengaruh Model Pembelajaran Kooperatif Think Pair Share (TPS) Menggunakan Assessment For Learning Pada Prestasi Siswa Sekolah Menengah Pertama,” Jurnal Karya Pendidikan Matematika, vol. 6, 2019, [Online]. Available: http://jurnal.unimus.ac.id/index.php/JPMat/index
4. N. Hanafi, S. Farmasari, M. Mahyuni, M. Amin, and Y. B. Lestari, “Pelatihan Pengembangan Model Penilaian Otentik (Authentic Assessment) pada Pembelajaran Bahasa Inggris Sekolah Dasar bagi Guru-Guru Bahasa Inggris Sekolah Dasar Di Kabupaten Lombok Barat,” Jurnal Pengabdian Magister Pendidikan IPA, vol. 4, no. 2, Jul. 2021, doi: 10.29303/jpmpi.v4i2.855.
5. G. Engelhard Jr and S. Wind, Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge, 2018.
6. S. A. Wind, “Examining the Impacts of Rater Effects in Performance Assessments,” Appl Psychol Meas, vol. 43, no. 2, pp. 159–171, Mar. 2019, doi: 10.1177/0146621618789391.
7. S. Sutrio, H. Sahidu, A. Harjono, I. W. Gunada, and H. Hikmawati, “Pelatihan Pembelajaran IPA Berbasis Inkuiri Berbantuan KIT Bagi Guru-Guru SD Di Kota Mataram,” Jurnal Pengabdian Masyarakat Sains Indonesia, vol. 2, no. 2, Nov. 2020, doi: 10.29303/jpmsi.v2i2.80.
8. D. Beutel, L. Adie, and M. Lloyd, “Assessment moderation in an Australian context: processes, practices, and challenges,” Teaching in Higher Education, vol. 22, no. 1, pp. 1–14, Jan. 2017, doi: 10.1080/13562517.2016.1213232.
9. J. Mason and L. D. Roberts, “Consensus moderation: the voices of expert academics,” Assess Eval High Educ, vol. 48, no. 7, pp. 926–937, 2023.
10. Y. Elizabeth Patras et al., “Meningkatkan Kualitas Pendidikan Melalui Kebijakan Manajemen Berbasis Sekolah Dan Tantangannya,” Jurnal Manajemen Pendidikan, vol. 7, no. 2, 2019.
11. D. Amir and A. Basit, “Kompetensi Pedagogik dan Profesional Mahasiswa Jurusan PAI pada Pelaksanaan PPL Tahun Akademik 2017/2018,” 2018.
12. S. V. Makwana, “Use of Grading System in Education,” RESEARCH REVIEW International Journal of Multidisciplinary, vol. 9, no. 4, pp. 260–266, Apr. 2024, doi: 10.31305/rrijm.2024.v09.n04.032.
13. T. A. Chowdhury, “International Journal of Linguistics and Translation Studies,” International Journal of Linguistics and Translation Studies, vol. 1, no. 1, pp. 32–1, 2020, doi: https://doi.org/10.36892/IJLTS.V1I1.14.
14. N. Abd Rahman, S. E. Mokshein, and H. Ahmad, “Validity and Reliability of Students’ Mathematical Process Rubric (Prom3) based on many-Facet Rasch Model (MRFM).,” International Journal of

Academic Research in Business and Social Sciences, vol. 11, no. 2, Feb. 2021, doi: 10.6007/ijarbss/v11-i2/9207.
15. N. Heidari, N. Ghanbari, and A. Abbasi, "Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating," Language Testing in Asia, vol. 12, no. 1, Dec. 2022, doi: 10.1186/s40468-022-00168-3.
16. Y. Du, B. D. Wright, and W. L. Brown, "Differential Facet Functioning Detection in Direct Writing Assessment," in Paper presented at the Annual Conference of the American Educational Research Association, Apr. 1996.
17. P. Sureeyatanapas, P. Sureeyatanapas, U. Panitanarak, J. Kraisriwattana, P. Sarootyanapat, and D. O'Connell, "The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&R) study: a case of English-speaking test," Language Testing in Asia, vol. 14, no. 1, Dec. 2024, doi: 10.1186/s40468-023-00271-z.
18. M. T. Braverman and M. E. Arnold, "An evaluator's balancing act: Making decisions about methodological rigor," New Dir Eval, vol. 2008, no. 120, pp. 71–86, 2008, doi: 10.1002/ev.277.
19. J. Liu and L. Xie, "Examining Rater Effects in a WDCT Pragmatics Test," Iranian Journal of Language Testing, vol. 4, no. 1, p. 50, 2014.
20. C. Myford and E. W. Wolfe, "Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II," 2004.
21. S. A. Wind and Y. Ge, "Detecting Rater Biases in Sparse Rater-Mediated Assessment Networks," Educ Psychol Meas, vol. 81, no. 5, pp. 996–1022, Oct. 2021, doi: 10.1177/0013164420988108.
22. T. G. Bond and C. M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press, 2013.
23. J. M. Linacre, "A user's guide to WINSTEPS® MINISTEP: Rasch-model computer programs," Chicago: MESA Press., p. 719, 2016.
24. G. Zhang, "Research on the application of multifaceted Rasch model analysis software facets in English test," in International Conference on Mechanisms and Robotics (ICMAR 2022), SPIE, 2022, pp. 760–765.
25. C. Elder, G. Barkhuizen, U. Knoch, and J. von Randow, "Evaluating rater responses to an online training program for L2 writing assessment," Language Testing, vol. 24, no. 1, pp. 37–64, Jan. 2007, doi: 10.1177/0265532207071511.
26. N. Khabbazbashi and E. D. Galaczi, "A comparison of holistic, analytic, and part marking models in speaking assessment," Language Testing, vol. 37, no. 3, pp. 333–360, Jul. 2020, doi: 10.1177/0265532219898635.
27. B. Sumintono, Aplikasi pemodelan rasch pada assessment pendidikan. Penerbit Trim Komunikata, 2015.
28. H. Y. Huang, "Modeling Rating Order Effects Under Item Response Theory Models for Rater-Mediated Assessments," Appl Psychol Meas, vol. 47, no. 4, pp. 312–327, Jun. 2023, doi: 10.1177/01466216231174566.
29. B. Sumintono and W. Widhiarso, Aplikasi model rasch: Untuk penelitian ilmu-ilmu sosial. Trim Komunikata Publishing House, 2013.
30. Linacre, "Standard errors: means, measures, origins and anchor values," Rasch Measurement Transactions, vol. 19, no. 3, p. 1030, 2005.
31. C. Van Zile-Tamsen, "Using Rasch Analysis to Inform Rating Scale Development," Res High Educ, vol. 58, no. 8, pp. 922–933, Dec. 2017, doi: 10.1007/s11162-017-9448-0.
32. H. Misbach and B. Sumintono, "Pengembangan dan Validasi Instrumen 'Persepsi Siswa tehadap Karakter Moral Guru' di Indonesia dengan Model Rasch 1," PROCEEDING Seminar Nasional Psikometri, vol. 148162, May 2014.
33. B. C. Wesolowski and S. A. Wind, "Pedagogical Considerations for Examining Rater Variability in Rater-Mediated Assessments: A Three-Model Framework," J Educ Meas, vol. 56, no. 3, pp. 521–546, Sep. 2019, doi: 10.1111/jedm.12224.
34. M. Wu, "Some IRT-based analyses for interpreting rater effects," 2017.
35. S. M. Wu and S. Tan, "Managing rater effects through the use of FACETS analysis: the case of a university placement test," Higher Education Research and Development, vol. 35, no. 2, pp. 380–394, Mar. 2016, doi: 10.1080/07294360.2015.1087381.

36. J. M. Linacre, A User's guide to WINSTEPS® Rasch-model computer programs: Program manual 4.4. 6. 2019.
37. D. Erguvan and B. Aksu Dunya, "Analyzing rater severity in a freshman composition course using many facet Rasch measurement," Language Testing in Asia, vol. 10, no. 1, Dec. 2020, doi: 10.1186/s40468-020-0098-3.
38. T. Eckes, "Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis," Lang Assess Q, vol. 2, no. 3, pp. 197–221, Oct. 2005, doi: 10.1207/s15434311laq0203_2.
39. T. Eckes, Introduction to many-facet Rasch measurement. Peter Lang, 2023.
40. C. Myford and E. W. Wolfe, "Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I," 2003. [Online]. Available: https://www.researchgate.net/publication/9069043
41. L. H. Asbulah, M. A. Lubis, A. Aladdin, and M. Sahrim, "KESAHAN DAN KEBOLEHPERCAYAAN INSTRUMEN PENGETAHUAN KOLOKASI BAHASA ARAB IPT (i-KAC IPT) MENGGUNAKAN MODEL PENGUKURAN RASCH," ASEAN Comparative Education Research Journal on Islam and Civilization (ACER-J), vol. 2, no. 1, pp. 97–106, 2018.
42. E. R. Lai, E. W. Wolfe, and D. H. Vickers, "Halo Effect s and Analytic Scoring: A Summary of Two Empirical Studies Research Report," 2012. [Online]. Available: http://www.pearsonassessments.com/research.
43. Y. A. Rahman, F. Apriyanti, and R. A. Nurdini, "Rater Severity/Leniency and Bias in EFL Students' Composition Using Many-Facet Rasch Measurement (MFRM)," Scope : Journal of English Language Teaching, vol. 8, no. 1, p. 258, Oct. 2023, doi: 10.30998/scope.v8i1.19432.
44. Z. Mohd Zabidi, B. Sumintono, and Z. Abdullah, "Enhancing analytic rigor in qualitative analysis: developing and testing code scheme using Many Facet Rasch Model," Qual Quant, vol. 56, no. 2, pp. 713–727, Apr. 2022, doi: 10.1007/s11135-021-01152-4.
45. R. Mohamat, B. Sumintono, and H. S. Abd Hamid, "Raters' Assessment Quality in Measuring Teachers' Competency in Classroom Assessment: Application of Many Facet Rasch Model," Asian Journal of Assessment in Teaching and Learning, vol. 12, no. 2, pp. 71–88, Nov. 2022, doi: 10.37134/ajatel.vol12.2.7.2022.