Fair or Fickle? The Tug-of-War Between Objectivity and Teacher-Student Bias in Speaking Assessments

Muhamad Firdaus Mohd Noh[1,*], Mohd Effendi Ewan Mohd Matore[2], and Nur Ainil Sulaiman[3]

[1] Faculty of Education, Universiti Kebangsaan Malaysia (UKM) Bangi, 43600, Selangor, Malaysia
[2] Research Centre of Leadership and Educational Policy, Faculty of Education, Universiti Kebangsaan Malaysia (UKM) Bangi, 43600, Selangor, Malaysia
[2] Centre of Teaching and Learning Innovation, Faculty of Education, Universiti Kebangsaan Malaysia (UKM) Bangi, 436000, Selangor, Malaysia
*p132669@siswa.ukm.edu.my

Abstract. A critical gap persists in understanding how student factors shape rating objectivity. This study examines the presence of biased interactions between teachers and students in the assessment of English language speaking performance. A total of 164 English teachers evaluated responses from 30 students across three item types (background interview, guided storytelling, and pair discussion) after receiving standardised training to enhance rating consistency. A linked rating design created systematic overlap among facets, enabling robust analysis with Many-Facet Rasch Measurement (MFRM). The preliminary MFRM calibration indicated acceptable model-data fit across the major facets, supporting the appropriateness of proceeding with the bias interaction analysis. Across 982 teacher × student interactions, 252 (25.66%) exceeded the two-tailed threshold of $|t| \geq 2.00$, indicating statistically significant bias; severity effects were slightly more common than leniency (53.17% vs. 46.83%). When disaggregated by student ability, mid-range students showed the largest share of biased interactions (45.63%), followed by high-ability (29.76%) and low-ability (24.60%) groups, suggesting that performances of ambiguous proficiency elicit greater rater variability, whereas very weak or strong performances are more readily distinguished. These patterns raise equity concerns, given that mid-range students constitute a substantial proportion of the population and appear more vulnerable to inconsistent ratings. The findings underscore the need for continuous professional development and structured, behaviourally anchored rubrics, complemented by routine bias audits, to mitigate rater effects and promote fairer evaluations. Limitations include a context-specific sample, a single assessment skill, and reliance on MFRM. Future research should ex-amine cross-subject and cross-level generalisability, use longitudinal designs, consider cultural/linguistic moderators, and rigorously evaluate bias-mitigation interventions and training.

Keywords: Rating Bias, Rater-Mediated Assessment, Multi-Facet Rasch Measurement, Speaking Test, Language Testing.

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_17

1 Introduction

The objectivity of educational assessments plays a crucial role in authentically capturing student performance. This concern is closely aligned with Sustainable Development Goal 4 (SDG 4), which emphasises inclusive and equitable quality education and the promotion of lifelong learning opportunities for all. SDG 4 places particular emphasis on reducing educational inequalities and ensuring that all learners have access to fair, meaningful, and quality learning opportunities. In this regard, fair assessment practices are essential because they help ensure that students are evaluated based on their actual performance rather than irrelevant sources of bias, inconsistent rating practices, or unequal assessment conditions [1]. Therefore, strengthening objectivity in assessment is not merely a technical psychometric concern, but also part of a broader educational commitment to equity, quality, and fairness. Psychometric principles dictate that assessments must be reliable, valid, and fair, ensuring they consistently yield the same results under comparable conditions, accurately measure the intended constructs, and remain unbiased across different demographic groups [1]. These principles underscore the necessity of standardised testing procedures, robust scoring methodologies, and meticulous interpretation of data to capture an authentic representation of performance [2]. Adherence to these standards enables assessment systems to deliver precise, equitable, and meaningful evaluations of students' competencies.

A robust assessment system begins with the careful selection and construction of test items, followed by the administration of the assessment, rating process, and the utilisation of results to impact students, educational institutions, and society [3]. In particular, the rating process is influential as it directly reflects how well a student's performance is evaluated according to established psychometric principles. Rating process stands as a pivotal facet in the assessment paradigm, as it exerts a profound impact on the appraisal of students' endeavours and guides the judgments made thereafter [4]. It involves assigning evaluative codes in the form of marks, grades, or statements that reflect a student's performance in the domains or skills being assessed. This process is essential for marking, assessing, and evaluating responses within a given domain based on pre-established standards. The approach to rating is influenced by the type of items used, the rating method, and the raters chosen by the administrators [5].

However, maintaining objectivity poses greater challenges, especially in the assessment of polytomous skills, as it depends heavily on raters' subjective judgments and often introduces unwarranted variability influenced by unintended factors [6]. Polytomous items are tasks that accommodate multiple levels of performance, requiring raters to assign scores across a range of categories rather than awarding a single holistic score. In language testing, such items typically involve productive tasks, such as speaking and writing skills, where candidates generate authentic language output which are then assessed using analytical criteria across several domains. Rater bias or interaction is among the four sources of rater variability, along with rater consistency, rater leniency or severity, and the rater's use of the rating scale [7]. Previous research

has highlighted the impact of rater-related variables, such as rating experience [8], rater training [5], and first language [9], on scoring consistency and fairness. The studies emphasise how subjective interpretations may affect the reliability of assessments involving open-ended responses. These biases can undermine the validity of the assessment process, leading to inconsistent and potentially unjust outcomes [10]. Despite these insights, limited attention has been given to how student-related factors influence rater behaviour and contribute to bias. Therefore, this study seeks to examine the presence of biased interactions between teachers and students in the assessment of English language speaking performance. The study is guided by one research question, "to what extent do teacher-student interactions influence bias in the marking of English language speaking proficiency?". The study hypothesises that there is no statistically significant bias interaction between teachers and students.

2 Literature Review

2.1 Conceptual Framework

This study's conceptual framework is anchored in the Lens Model [11], Social Constructivism Theory [12] and the Social Constructionism Theory [13] to provide a theoretically robust explanation of how bias interactions emerge in speaking assessments. The Lens Model conceptualises assessment as an inferential process in which raters do not directly observe the latent construct (in this study is speaking proficiency) but instead rely on multiple observable cues to make judgments. In speaking assessments, these cues include student linguistic performance, paralinguistic features, task characteristics, scoring domains, and rater-related characteristics [14]. The model emphasises that judgment accuracy depends on how consistently and appropriately raters utilise these cues relative to the construct being assessed [15]. Misalignment between the ecological validity of cues (how strongly they represent speaking proficiency) and rater cue utilisation can result in distorted judgments, thereby opening space for systematic bias. Within this framework, bias is understood as a product of cue interactions rather than isolated rater error [16]. The interaction between teacher-related cues, such as expectations, beliefs and professional identity, and student-related cues, including perceived ability, confidence, accent, or interactional style, is particularly salient in speaking assessments, which are inherently subjective and socially mediated. The Lens Model provides an analytic structure for examining how such interactions influence scoring outcomes, including tendencies toward severity or leniency, differential treatment of students, and construct-irrelevant variance [17]. By explicitly modelling these cue–judgment relationships, the framework enables a meticulous investigation of how bias is produced, maintained, and manifested in final speaking scores.

Social Constructivism Theory complements the Lens Model by foregrounding the socio-cognitive processes through which raters interpret and apply assessment criteria [12]. From a constructivist perspective, teachers' rating behaviours are shaped by accumulated experiences, professional socialisation, and shared assessment practices

within their communities. Prior knowledge gained through rater training, alongside experiential factors such as years of teaching, exposure to diverse learners, and repeated assessment encounters, informs how raters make sense of performance descriptors and apply scoring standards. This theoretical lens underscores that bias is not merely an individual cognitive flaw but a socially situated phenomenon, emerging from the interaction between raters' internalised knowledge structures and the immediate assessment context. The Lens Model and Social Constructivism Theory offer a complementary and integrated foundation for examining bias interactions in English language speaking assessments.

The social construction of reality [13] view further strengthens this theoretical positioning by explaining that knowledge, meaning, and judgment are not formed in isolation, but are socially constructed through interaction, institutional routines, and shared interpretive practices. In the context of speaking assessment, teachers' rating decisions may therefore reflect not only individual cognitive judgment but also the assessment norms, professional expectations, and collective understandings developed within their educational community. This perspective supports the argument that bias interactions can emerge when raters interpret student performance through socially shaped assumptions about ability, fluency, confidence, or acceptable language use. Thus, Berger and Luckmann's [13] notion of socially constructed reality provides an additional theoretical basis for understanding rating bias as a socially embedded phenomenon rather than merely a technical scoring error.

2.2 Rating in Language Assessment

In the context of language assessment, rating takes on unique characteristics due to the nature of the skills being evaluated [18]. Language assessments often include both objective items, such as multiple-choice questions, and subjective items, such as essay writing. Objective items are typically graded using a dichotomous method, where students receive either a full mark or no mark for each response [19]. This type of rating is straightforward, with raters relying on answer keys that provide the correct responses. Subjective items, on the other hand, require students to construct their own answers without being presented with a set of options [5]. These items demand a higher level of cognitive engagement from the students and are evaluated based on a broader range of acceptable responses. In language assessments, subjective rating is often guided by rubrics, which outline the criteria and quality definitions across various performance levels [20, 21]. Rubrics are crucial for ensuring that complex language skills, such as speaking or writing, are assessed comprehensively, considering factors like vocabulary usage, grammatical accuracy, clarity of expression, fluency, and interaction capabilities [22].

The rating process in language assessments can be conducted holistically or analytically. Holistic rating involves assigning a single overall score that reflects the students' general ability based on their performance, which is efficient but less detailed [23]. Analytical rating, on the other hand, involves assessing multiple specific criteria

or domains, providing a richer, more diagnostic insight into the students' strengths and weaknesses [24]. While holistic rating saves time and resources, analytical rating offers more detailed feedback that can be critical for guiding future learning and instructional strategies [25]. The effectiveness of rating in language assessment depends on the careful preparation and application of an appropriate rubric that is well-structured, clearly defines performance criteria, and includes specific descriptors for different levels of achievement, ensuring that raters can consistently and accurately evaluate various language skills. [26]. The rating system must be meticulously designed and implemented to ensure it meets the desired quality standards, providing valid and reliable outcomes that accurately reflect the students' language abilities. Recent research in language assessment has increasingly focused on the improvement of rater reliability and rating designs to address challenges associated with subjectivity in the evaluation of complex skills, such as speaking and writing skills [27].

2.3 Bias in the Rating Process

Bias in the rating process refers to the systematic deviation in the scores assigned by raters that is not attributable to the student's actual performance [28]. This bias can emerge from various sources pertaining to the assessment system, including the raters themselves, students, items, domains and assessment setting [29]. The rating process, which is essential for evaluating a student's performance, can be significantly impacted by such biases, leading to inaccurate and unfair results. Inaccurate results occur when biases, such as leniency or severity, cause raters to either overestimate or underestimate a student's actual performance, leading to scores that do not accurately reflect the student's true abilities [20, 22]. Unfair results arise when inconsistencies in rating, such inclination to assess differently based on factors coming from the raters, students or the assessment setting which result in unequal treatment of students, potentially disadvantaging certain individuals or groups despite their actual performance [30]. Bias may arise from raters' subjective tendencies, such as being more severe or more lenient with certain students, items, or domains, as well as differences in their levels of experience, ultimately affecting the reliability and validity of the assessment outcomes [6]. The incident of biased items is evident when test questions or tasks are designed in a manner that favours certain groups of students over others, often due to gender, cultural, linguistic, or socioeconomic factors [31]. Such bias creates unequal opportunities for students to showcase their true abilities, skewing the results in favour of those whose backgrounds align more closely with the content of the items [32].

Understanding the sources and impact of bias is crucial for ensuring that assessments are fair and accurately reflect a student's performance. By examining the interplay of these biases through theoretical models like the lens model and employing analytical tools such as the Many-Facet Rasch Model (MFRM), educators and researchers can better identify and address factors that compromise the reliability and validity of assessment ratings [33]. The lens model developed by Brunswik [11] posits that there are intercorrelations among the signals within the rating system, meaning that interactions between different elements, such as the raters, students, items, and

domains, can influence the rating process. The lens model has been employed by previous researchers in other fields of subjective assessment, such as creativity assessment, communication and peer assessment of writing skill, as a theoretical framework [15, 34, 35]. The importance of analysing these biases lies in ensuring that the diversity among raters does not lead to irrelevant construct variance, which could undermine the fairness and accuracy of the rating process [36]. Meanwhile, biases in rating can be identified through qualitative observations or quantitative statistical analysis using models like MFRM model, which helps determine the extent to which bias affects the outcomes [30].

MFRM model has proven highly useful in previous research, particularly in addressing issues related to rater effects and bias in performance assessments, as it allows researchers to analyse multiple facets, such as rater severity, leniency, and inconsistencies in scoring [19, 35]. The model also enables the investigation of interactions between various factors, such as the test-taker's abilities, rater backgrounds, and task difficulty, offering a more comprehensive understanding of the rating process and providing insights into how different variables affect scores [37]. Additionally, MFRM has been widely used in language assessment research to evaluate the effectiveness of rater training and monitor the consistency of raters over time [38].

2.4 Bias Interaction Between Teachers and Students

One of the most frequently studied forms of bias in rating is the interaction between teachers and students. Previous research has shown that teachers may exhibit bias based on various student characteristics, such as their academic performance, first language, individual traits, and age [9, 39–42]. For instance, Bijani and Khabiri [42] reported that teachers tend to show more bias towards students at the extremes of the ability spectrum, either very high or very low achievers [42]. Bias is often more pronounced when assessing low-achieving students, as their responses may lack clarity, making it challenging for teachers to maintain consistent rating standards [43]. Additionally, teachers may exhibit bias based on their familiarity with the student's first language. For example, raters who share a linguistic background with the student may grade them more leniently, which can lead to a tendency for raters to award higher marks to lower-achieving students, placing them on the same level as their higher-achieving peers [44, 45].

Bias can also occur due to the student's age, with younger students often receiving more variable scores, especially if their responses do not meet certain proficiency standards [46]. These biases underscore the need for rigorous analysis and mitigation strategies to ensure that all students are assessed fairly, regardless of their background or abilities. The interaction between teachers and specific test items is another critical area where bias can emerge. While research on this type of bias is less extensive, it is crucial given that language assessments often use multiple items to evaluate different domains or constructs. This variety can introduce opportunities for bias, particularly if raters have different levels of comfort or familiarity with certain items [47]. For

example, teachers may exhibit bias when rating different genres of writing, such as narrative versus expository essays, where they might apply stricter standards to one genre over the other [48]. Studies have shown that even when teachers are trained to assess multiple item types, bias can still occur, particularly when the items require assessing different skills or knowledge areas. Identifying and addressing these biases is essential to ensure that assessments accurately reflect the student's abilities across all items, without undue influence from the rater's preferences or biases.

2.5 Assessment Practices in Malaysia

In Malaysia, the practice of assessment, including rating, is predominantly school-based, with teachers playing a central role in evaluating student performance from primary through secondary school [49]. Teachers are responsible for designing a variety of assessments, such as quizzes, projects, oral presentations, and summative exams, tailored to the needs of their students [50]. As the primary raters, they are tasked with assigning grades based on their evaluations, which requires them to uphold objectivity and fairness while managing a significant workload. The lack of external, standardised testing during most of the school years places immense responsibility on teachers to ensure that their assessments align with national curriculum standards and reflect the true capabilities of students. In particular, subjective assessments, such as those related to language skills and arts, demand careful consideration to avoid bias in rating. To support the integrity and consistency of these teacher-based assessments, Malaysia employs a system of mentoring and monitoring [51]. Mentoring provides teachers with guidance from more experienced educators or subject experts, helping them refine their assessment and rating techniques [52]. Meanwhile, monitoring is conducted by the ministry or local education officers, who monitor the implementation of assessment practices through school visits, reviews of student work, moderation, and inspections of rating records [53]. Despite the central role teachers play in assessment, assessment literacy among teachers in Malaysia remains relatively low, particularly in areas such as designing valid assessments, applying consistent rating criteria, administering, scoring, and interpreting assessment results [54].

3 Methods

3.1 Samples

The study employed a multistage cluster sampling design to recruit 164 English language teachers from Selangor, Malaysia, representing a diverse range of school types and teaching contexts. All participants held at least a bachelor's degree and taught English to lower-secondary students (Forms 1–3). They reported varied professional backgrounds in classroom teaching, student performance rating, and participation in rater training or professional development related to language pedagogy and assessment. To capture this variation, teachers were categorised along three dimensions: teaching experience, rating experience, and prior rater training experience (see Table 1). Teaching experience was grouped into 1 to 10 years, 11 to 20 years, and

more than 20 years. Rating experience was grouped into no experience, 1 to three 3, and more than 3 years. Prior rater training was coded as attended or not attended. These variables were included because they represent theoretically meaningful differences in pedagogical exposure, assessment literacy, and familiarity with scoring standards.

Table 1. Teachers' experience profiles (n = 164)

Experiences	Frequency	Percentage
Teaching experience		
1 to 10 years	50	30.5
11 to 20 years	56	34.1
More than 20 years	58	35.4
Rating experience		
No experience	63	38.4
1 to 3 years	44	26.8
More than 3 years	57	34.8
Training experience		
Attended rater training	102	62.2
Never attended rater training	62	37.8

3.2 Instruments

The instruments comprised audio-recorded responses from 30 students who completed an English language speaking test, an analytical scoring rubric, a scoring sheet, and detailed scoring guidelines. The speaking test consisted of three item types adapted from the Form Three English Language Speaking Test Bank: Item 1, a background interview; Item 2, a guided storytelling task; and Item 3, a pair discussion task. The exact interlocutor instructions, visual stimuli, and back-up prompts are provided in Appendix A [55]. Item 1 required candidates to answer personal and familiar-topic questions, including questions about their name, place of residence, English lessons, homework routines, favourite song, and activities with friends. Item 2 required each candidate to narrate a story based on a sequence of five pictures. This item included two parallel visual prompts, one for Candidate A and one for Candidate B. Item 3 required paired candidates to discuss a school jungle-trekking activity and decide which item students should bring would be the most important. Back-up prompts were provided for Item 3, but these were used only when necessary to support interaction and were not treated as independent test prompts. For the main MFRM analysis, the three speaking items were treated as the item type facet. Prompt was not included as a separate facet because the assessment design did not systematically cross multiple prompts within each question type across all students and raters.

The analytical scoring rubric consisted of three criteria: grammar, vocabulary, and communicative competence. Each criterion was rated on a 0–5 scale. For grammar, the rubric assessed control of simple grammatical structures, sentence formation, and use of basic cohesive devices. For vocabulary, it assessed range, appropriateness, relevance, and repetition in discussions of everyday situations and familiar topics. For communicative competence, it assessed clarity, hesitation, the ability to initiate,

maintain, develop, and close a conversation, and the ability to ask for clarification or signal a misunderstanding. The full rubric is provided in Appendix B. The rubric was deemed usable for the present study because it provided analytically distinct scoring criteria aligned with the construct of speaking proficiency. The descriptors were behaviourally anchored at Bands 5, 3, 1, and 0, allowing raters to distinguish high, moderate, limited, and insufficient performance. Bands 4 and 2 functioned as intermediate categories between adjacent descriptor levels. To reduce ambiguity in interpreting these intermediate bands, the rubric was accompanied by scoring guidelines and rater training before rating began. The quality of the scale was also examined empirically through rating-scale functioning analysis. All six score categories were used more than ten times, the average measures increased monotonically across the categories, the category fit statistics were acceptable, and the thresholds advanced in the expected order. These results supported the usability of the 0–5 rating scale for the speaking assessment.

Content validity was evaluated to ensure that the speaking tasks, analytical rubric, and scoring guidelines adequately represented the targeted construct of English language speaking proficiency and were appropriate for Form Three learners. Five expert panelists, each holding a doctorate in Teaching English as a Second Language (TESL), independently reviewed the assessment components. The Content Validity Ratio (CVR) was calculated using Lawshe's method [56] to assess whether each component was essential for assessing speaking proficiency, as determined by five expert panels. The CVR analysis covered the three item types, the three rubric criteria, and the scoring guidelines. All evaluated components achieved acceptable CVR values and were therefore retained. The CVR results are reported in Table 2.

Table 2. Content validity ratio (CVR) results for the speaking assessment components

Assessment component	Component evaluated	CVR value	Decision
Speaking items	Item 1: Background interview	1.00	Retained
	Item 2: Guided storytelling	1.00	Retained
	Item 3: Pair discussion	1.00	Retained
Rubric criterion	Grammar	1.00	Retained
	Vocabulary	1.00	Retained
	Communicative competence	1.00	Retained
Scoring document	Scoring sheet	1.00	Retained
	Scoring guidelines	1.00	Retained

Note. A CVR value of 1.00 indicates full expert agreement that the component was essential for assessing Form Three English speaking proficiency.

3.3 Data Collection Procedures

All participating teachers received standardised rater training before scoring the student recordings. The training was designed to promote a shared understanding of the

speaking construct, the analytical rubric, the scoring categories, and the rating procedures. It covered the structure of the speaking test, the three item, the three rubric criteria, the 0–5 score categories, and the use of the scoring sheet. Teachers were also familiarised with the audio-recorded responses and the scoring guidelines before the formal rating task began. The training was intended to reduce unnecessary subjectivity and improve consistency in applying the rubric. However, the study did not treat training as a bias-removal mechanism. Rather, training was used as a standardisation procedure before examining whether meaningful rater effects and teacher–student bias interactions still occurred. This distinction is important because the study aimed to investigate bias in a realistic rater-mediated speaking assessment context, where rater training may reduce but not eliminate severity, leniency, or differential rater functioning. A linked rating design was implemented to create systematic overlap across the assessment facets. The main facets were teachers, students, item, and scoring domains. Each student was independently rated by approximately ten teachers, and each teacher scored a random subset of approximately six students. This partially balanced structure yielded 6,886 score units and enabled the MFRM analysis to estimate rater severity, student ability, task difficulty, domain difficulty, and selected interaction effects on a common logit scale.

3.4 Data Analysis Procedures

The data were analysed using MFRM in FACETS 3.87.0. The first stage of analysis calibrated the main assessment facets: teachers, students, item types, and scoring domains. The teacher facet was interpreted as rater severity or leniency, the student facet as speaking ability, the item facet as item difficulty, and the domain facet as criterion difficulty. Teacher background variables, namely rating experience, prior rater training, and teaching experience, were also included as grouping variables to support interpretation of rater-related patterns. Model-data fit was examined before interpreting bias interactions. Item fit was evaluated using infit and outfit mean-square statistics, standardised fit values, standard error, and point-measure correlations. Rater fit was examined using infit and outfit mean-square statistics to determine whether teachers applied the scoring scale consistently across students and items. In line with common MFRM interpretation for rater-mediated assessment, rater fit values between 0.50 and 1.50 were interpreted as productive fit, values greater than 1.50 as underfit or inconsistent rating behaviour, and values below 0.50 as overfit or overly predictable rating behaviour. These fit analyses were conducted before the bias analysis to ensure that the interpretation of teacher–student interactions was grounded in an acceptable measurement model.

Rating-scale functioning was also examined to assess the quality and usability of the 0–5 rubric. The analysis considered category frequency, monotonic advancement of average measures, category fit, ordered Rasch-Andrich thresholds, threshold distance, and category probability curves. The purpose was to determine whether raters used the full score scale meaningfully and whether the categories functioned in the intended order. Differential rater functioning was examined through bias interaction analysis.

The primary analysis focused on teacher × student interactions because the study's main objective was to determine whether teachers showed statistically significant severity or leniency toward specific students after accounting for other facets of the measurement system. Bias contrast measures were interpreted on the logit scale. Positive bias measures indicated relative severity, meaning the teacher scored a student more harshly than the model expected. Negative bias measures indicated relative leniency, meaning that the teacher scored a student more generously than expected. Bias interactions were flagged as statistically meaningful when the absolute t-value was greater than or equal to 2.00, corresponding to a two-tailed significance level of approximately $p < .05$.

To address whether rater behaviour differed according to item type, additional interaction evidence was considered for teacher × item. This allowed the analysis to examine whether raters were more severe or more lenient when rating the background interview, guided storytelling, or pair discussion task. Student ability groups were created using student logit measures to support the descriptive examination of whether teacher–student bias interactions were concentrated among high-, mid-, or low-ability students. This grouping approach follows previous MFRM-based bias studies that examined interaction patterns by classifying examinees according to their logit-based ability locations [44, 57, 58]. In the present study, students with logit measures from 1.00 to 2.82 were classified as high ability, students with logit measures from -0.99 to 0.99 were classified as mid-range ability, and students with logit measures from -2.12 to -1.00 were classified as low ability. These cut-offs were used descriptively rather than as externally standardised proficiency thresholds, with the aim of comparing the distribution of significant teacher–student bias interactions across different regions of the speaking-ability continuum. The mid-range group was theoretically important because performances near the middle of an analytic rating scale are often less clearly distinguishable than clearly high or clearly low performances, which increases raters' reliance on personal weighting of criteria, implicit thresholds, and adjacent score-band decisions [59, 60].

4 Results

4.1 Preliminary Model-Data Fit

Before examining bias interactions, the measurement quality of the speaking assessment was inspected through item type fit, item separation, rating-scale functioning, rater severity, and rater fit. The MFRM analysis calibrated four main facets: teachers, students, item types, and scoring domains. Table 3 presents the item measures and fit statistics. The item measures ranged from -0.20 logits to 0.33 logits. Pair discussion was the most difficult item, with a measure of 0.33 logits, followed by guided storytelling at -0.13 logits and background interview at -0.20 logits. The small logit spread indicated that the three speaking items were relatively balanced in difficulty. All three items showed acceptable fit. The infit MnSq values ranged from 1.00 to 1.01, while the outfit MnSq values ranged from 0.99 to 1.01. The standardised

fit values were close to zero, ranging from -0.3 to 0.1. The standard errors for all three items were 0.03, indicating precise item estimates. The PTMEA correlations were positive and high, ranging from .69 to .73. These results indicate that all three speaking items functioned productively within the measurement model.

Table 3. Item measures and fit statistics

Item	Total score	Logits	S.E.	Infit		Outfit		PTMEA	
				MnSq	Zstd	MnSq	Zstd	Corr.	Exp.
Background interview	5195	-0.20	0.03	1.01	0.1	1.01	0.1	0.71	0.71
Guided storytelling	5120	-0.13	0.03	1.00	-0.1	0.99	-0.1	0.69	0.71
Pair discussion	7126	0.33	0.03	1.00	-0.1	0.99	-0.3	0.73	0.71
Mean	5813.7	0.00	0.03	1.00	0.0	1.00	-0.1	0.71	-
SD (Population)	928.5	0.23	0.00	0.00	0.2	0.01	0.2	0.01	-
SD (Sample)	1137.1	0.29	0.00	0.01	0.2	0.01	0.3	0.02	-

The item separation statistics further supported the quality of the item facet (see Table 4). The item separation ratio was 9.01, the item separation index was 12.35, and the item separation reliability was .99. These results indicate that the three item types were estimated with high precision and were sufficiently differentiated along the item-difficulty continuum.

Table 4. Item separation statistics

Statistic	Item facet
Separation ratio	9.01
Separation index	12.35

4.2 Rating-Scale Functioning of the Analytic Rubric

The functioning of the 0-5 scoring scale was examined to determine whether the analytic rubric operated as intended. The six score categories were all used more than ten times. As illustrated in Table 5. Category use ranged from 86 observations for score 0 to 2711 observations for score 3. The observed average measures increased monotonically from -2.12 logits for score 0 to 2.94 logits for score 5. The outfit MnSq

values ranged from 0.9 to 1.3, indicating acceptable category-level fit. The Rasch-Andrich thresholds also advanced in the expected order from -4.27 to 3.98. These results indicate that the 0–5 rating scale functioned appropriately and that all score categories were retained.

Table 5. Rating-scale functioning of the 0–5 analytic rubric

Score category	Frequency used	Percentage	Average measure	Outfit MnSq	Rasch Andrich threshold
0	86	1%	-2.12	1.3	-
1	807	12%	-1.63	0.9	-4.27
2	2278	33%	-0.52	1.0	-2.10
3	2711	40%	0.66	1.0	-0.12
4	785	11%	1.87	1.0	2.51

The threshold distances between adjacent score categories were also examined. All threshold distances fell within the acceptable range of 1.0 to 5.0 logits, except for the structural transition from score 0 to score 1, where score 0 functioned as the lowest category (see Table 6). No disordered threshold pattern was observed. Therefore, no score category was collapsed or removed.

Table 6. Threshold distance between adjacent score categories

Adjacent categories	Threshold distance	Decision
0–1	4.27	Retained
1–2	2.17	Retained
2–3	1.98	Retained
3–4	2.72	Retained
4–5	1.47	Retained

4.3 Rater Severity and Rater Fit

The teacher facet was examined to identify rater severity and leniency before interpreting bias interactions. Teacher measures ranged from 2.59 logits for the most severe teacher to -2.81 logits for the most lenient teacher (see Table 7). The standard errors for teacher measures were low, ranging from 0.22 to 0.26, indicating precise estimates of teacher severity. Teachers were grouped into three severity categories based on their logit measures. A total of 27 teachers were classified as severe raters, 111 as within the acceptable severity range, and 26 as lenient raters. Thus, most teachers were located within the accepted severity range.

Table 7. Rater severity categories

Rater severity category	Logit criterion	N	%
Severe raters	> 1.00 logits	27	16.5
Accepted severity range	-1.00 to 1.00 logits	111	67.7
Lenient raters	< -1.00 logits	26	15.9
Total	-	164	100.0

Rater fit was also examined to determine whether teachers applied the rating scale consistently. A total of 112 teachers showed productive fit, 27 teachers showed underfit or inconsistent rating behaviour, and 25 teachers showed overfit or overly predictable rating behaviour (see Table 8). These findings indicate that most teachers demonstrated acceptable fit to the MFRM model, while a smaller group showed rating patterns requiring further inspection.

Table 8. Rater fit categories

Rater fit category	Fit interpretation	N	%
Productive fit	Acceptable rating consistency	112	68.3
Underfit	Inconsistent rating pattern	27	16.5
Overfit	Overly predictable rating pattern	25	15.2
Total		164	100.0

4.4 Rater Background Variables

The rater background variables were examined to determine whether severity range and rater fit varied by rating experience, rater training experience, and teaching experience. As shown in Table 10, the widest severity range for rating experience was observed among teachers with four to six years of rating experience (5.40 logits), followed by those with one to three years of rating experience (4.47 logits) and those with no rating experience (4.00 logits). For teaching experience, teachers with 11 to 20 years of experience showed the widest severity range (4.75 logits), followed by those with 1 to 10 years of experience (4.60 logits) and those with more than 20 years of experience (4.34 logits). For rater training experience, teachers who had attended rater training showed a wider severity range (5.40 logits) than those who had not (4.80 logits).

Table 9. Rater severity range and fit by rater background variables

Background variable	Group	Logits	Minimum logits	Maximum logits	Logit range	Underfit raters (%)	Overfit raters (%)
Rating experience	No experience	0.03	-1.94	2.06	4.00	17.5	14.3
	1 to 3 years	-0.06	-2.74	1.73	4.47	22.7	22.7
	4 to 6 years	-0.20	-2.81	2.59	5.40	7.0	14.0
Teaching experience	1 to 10 years	0.02	-2.81	1.79	4.60	18.0	12.0
	11 to 20 years	-0.07	-2.74	2.01	4.75	17.9	14.3
	More than 20 years	-0.10	-1.75	2.59	4.34	10.3	22.4
Rater training experience	Attended rater training	-0.01	-2.81	2.59	5.40	10.8	17.6
	Did not attend rater training	-0.07	-2.74	2.06	4.80	22.6	14.5

The fixed chi-square results are presented in Table 10. The analysis showed a statistically significant difference in rater severity across rating-experience groups, $\chi^2(2) = 26.0$, $p < .05$. A statistically significant difference was also found across teaching-experience groups, $\chi^2(2) = 7.3$, $p = .02$. However, the difference between teachers who had attended rater training and those who had not attended rater training was not statistically significant, $\chi^2(1) = 3.1$, $p = .08$.

Table 10. Fixed chi-square results for rater background variables

Rater background variable	df	Chi-square	p-value	Decision
Rating experience	2	26.0	< .05	Significant
Rater training experience	1	3.1	.08	Not significant
Teaching experience	2	7.3	.02	Significant

4.5 Differential Rater Functioning by Item

The interaction between teachers and item types was examined to determine whether teachers showed differential severity or leniency across the three speaking tasks. A total of 492 teacher × item interactions were examined, based on 164 teachers and three task types. Overall, 33 significant teacher × item bias interactions were identified,

representing 6.71% of all teacher × item interactions (see Table 12). These significant interactions involved 24 teachers, or 14.63% of the full rater sample. The highest number of item-bias interactions occurred in the background interview, with 13 significant interactions, representing 7.93% of all teacher × background interview interactions. This was followed by pair discussion, with 12 significant interactions (7.32%), and guided storytelling, with 8 significant interactions (4.88%). The direction of task-type bias varied by task. In the pair discussion, 11 of the 12 significant interactions indicated leniency, while 1 indicated severity. For guided storytelling, seven of the eight significant interactions indicated severity, while one indicated leniency. For the background interview, eight interactions indicated severity and five indicated leniency.

Table 11. Teacher × item bias interactions

Item	Item measure	Interactions		Significant interactions	
		Severity	Leniency	N	%
1	0.33	1	11	12	7.32
2	-0.13	7	1	8	4.88
3	-0.20	8	5	13	7.93
Total	-	16	17	33	6.71

4.6 Teacher–Student Bias Interactions

The main bias analysis examined teacher × student interactions. A total of 982 teacher × student interactions were analysed. The largest positive bias measure was 4.10 logits, observed for the interaction between Teacher 75 and Student 24. The largest negative bias measure was -4.84 logits, observed for the interaction between Teacher 126 and Student 2. Based on the direction of the bias measure, positive values indicated leniency and negative values indicated severity. Overall, 253 out of 982 teacher × student interactions exceeded the significance criterion of $|t| \geq 2.00$. This represents 25.76% of all teacher × student interactions. The remaining 729 interactions, or 74.24%, did not exceed the significance criterion. Among the significant teacher × student interactions, 134 were severity interactions and 118 were leniency interactions (see Table 12).

Table 12. Summary of teacher × student bias interactions

Statistic	Teacher × student interactions
Total teacher × student interactions	982
Maximum positive bias measure	4.10
Maximum negative bias measure	-4.84
Significant bias interactions,	252 (25.66%)
Non-significant interactions	729 (74.24%)
Severity interactions	134 (53.17%)
Leniency interactions	118 (46.83%)

Figure 1 presents the distribution of teacher × student bias t-values across student logit measures. The horizontal reference lines mark the upper and lower control limits at +2.00 and -2.00. Interactions located beyond these limits were classified as statistically significant bias interactions.

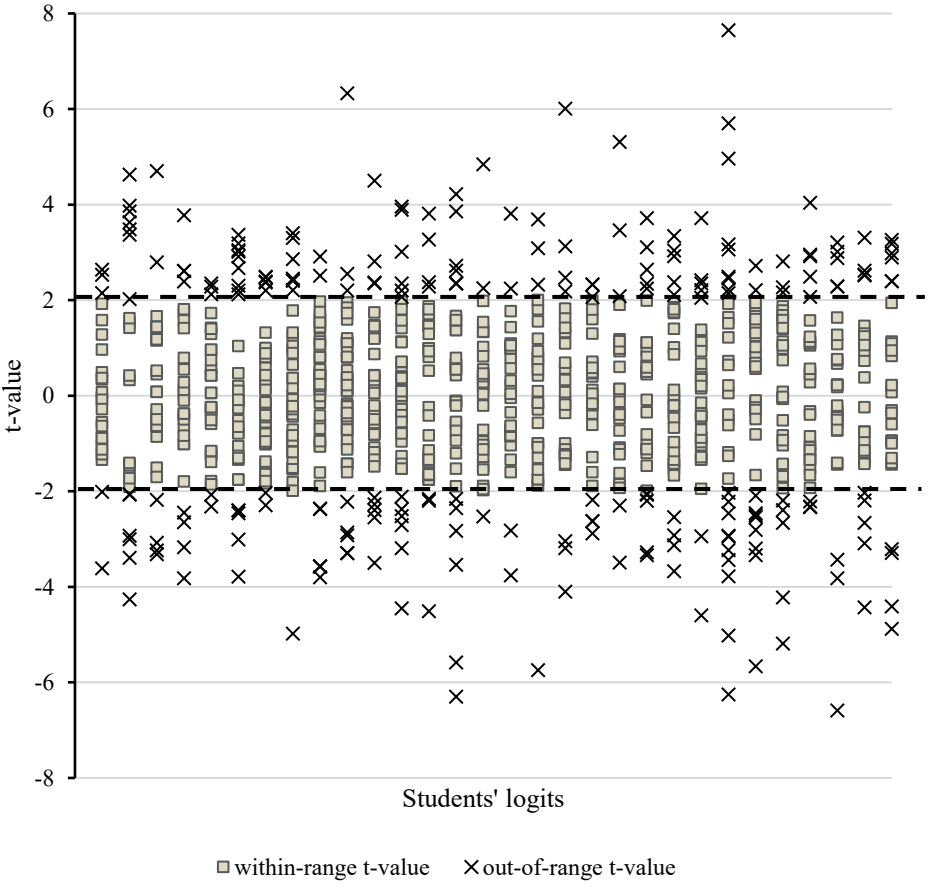


Fig. 1. Distribution of t-value for the bias interaction between teachers and students

The percentage of significant bias interactions was interpreted as the prevalence of bias, whereas the bias measure in logits was interpreted as the magnitude of bias. The t-value was used only as the statistical decision criterion, with interactions exceeding $|t| \geq 2.00$ classified as statistically significant. To describe effect size, the logit bias measures were inspected for all significant teacher × student interactions. Positive bias measures indicated unexpectedly lenient scoring, whereas negative bias measures indicated unexpectedly severe scoring. The largest positive bias measure was 4.10 logits, while the largest negative bias measure was -4.84 logits, indicating that some

teacher × student interactions involved substantial deviations from model-expected ratings.

The teacher × student bias interactions were further examined according to rater severity profile (see Table 13). Teachers in the accepted severity range showed the highest number of significant teacher × student bias interactions, with 177 out of 664 interactions, representing 26.66% (see Table 14). Severe raters showed 39 significant interactions out of 162 interactions, representing 24.07%. Lenient raters showed 37 significant interactions out of 156, representing 23.72%.

Table 13. Teacher × student bias by rater severity profile

Rater severity profile	N teachers with bias	N interactions	Significant bias interactions	
			N	%
Severe raters	21	162	39	24.07
Lenient raters	17	156	37	23.72
Accepted severity range	85	664	177	26.66

Teacher × student bias interactions were also examined according to rater fit profile. Underfitting raters showed the highest proportion of significant teacher × student bias interactions, with 75 significant interactions out of 162 interactions, representing 46.30% (see Table 14). Productive-fit raters showed 171 significant interactions out of 670 interactions, representing 25.52%. Overfitting raters showed seven significant interactions out of 150, representing 4.67%.

Table 14. Teacher × student bias by rater fit profile

Rater fit profile	N teachers with bias	N interactions	Significant bias interactions	
			N	%
Underfit raters	26	162	75	46.30
Overfit raters	4	150	7	4.67
Productive-fit raters	93	670	171	25.52

The significant teacher × student bias interactions were also analysed according to student ability group. Students were grouped into high-, mid-range-, and low-ability groups based on their logit measures. High-ability students were those with measures from 1.00 to 2.82 logits (see Table 15). Mid-range-ability students were those with measures from -0.99 to 0.99 logits. Low-ability students were those with measures from -2.12 to -1.00 logits. The mid-range-ability group had the highest number of significant bias interactions, with 115 interactions, representing 45.63% of the ability-group bias total. The high-ability group had 75 significant interactions, representing 29.76%. The low-ability group had 62 significant interactions, representing 24.60%.

Table 15. Teacher × student bias interactions by student ability group

Student ability group	Logit range	N students	N interactions	Significant bias interactions			
				N	%	Severity	Leniency
High ability	1.00 to 2.82	8	246	75	29.76	39	36
Mid-range ability	-0.99 to 0.99	16	531	115	45.63	67	48
Low ability	-2.12 to -1.00	6	205	62	24.60	28	34
Total	-	30	982	252	100.00	134	118

5 Discussion

The three speaking items showed acceptable item type fit, low standard errors, and positive PTMEA correlations. These results indicate that the background interview, guided storytelling and pair discussion items functioned productively within the MFRM model. The item separation statistics further suggested that the item facet was estimated with high precision. Bias interpretation should proceed only after the measurement system has demonstrated acceptable fit to the model. In this study, the item fit evidence suggests that the observed bias patterns were not simply artefacts of poorly functioning speaking items. The analysis therefore supports the use of the calibrated MFRM estimates as a basis for interpreting item difficulty, rater severity, student ability, domain difficulty, and interaction bias [61–63].

The rating-scale analysis also supports the usability of the analytic rubric. All six score categories were used more than ten times, the average measures increased monotonically, the outfit mean-square values were acceptable, and the Rasch-Andrich thresholds advanced in the expected order. These findings indicate that the 0–5 scoring scale was not functioning randomly or redundantly. Instead, raters were generally able to use the categories in a meaningful order when evaluating grammar, vocabulary, and communicative competence. This strengthens the argument that the rubric was empirically usable, not merely theoretically relevant. The intermediate categories also appeared to function adequately, although their interpretation still depended on rater understanding of adjacent band descriptors. For this reason, the evidence supports retaining the full scale, while also highlighting the need for rater training and calibration to ensure shared interpretation of the middle score bands [64, 65].

Within this acceptable measurement structure, the teacher × student bias analysis showed that 25.66% of interactions exceeded the statistical criterion of $|t| \geq 2.00$. The present findings substantiate the concern that rater bias is a widespread phenomenon in a subjective assessment such as speaking skills. Over one-fourth of the teacher–student interactions exhibited statistically significant bias, which reaffirms that subjectivity remains a formidable challenge in the assessment of polytomous language tasks [14]. This is consistent with previous studies indicating that, despite the use of analytic

rubrics and standardised training, rater-mediated assessments remain susceptible to inconsistencies arising from individual interpretations of performance features [66]. In speaking assessment, raters are required to make rapid, integrative judgments of multiple linguistic and interactional dimensions simultaneously, often under time constraints and without the opportunity for repeated observation [5]. Although analytic rubrics attempt to decompose performance into discrete criteria, raters must still infer underlying proficiency from surface features of speech, which may be interpreted differently depending on personal benchmarks and prior assessment experiences [67]. Speaking assessments, in particular, impose a high cognitive load because constructs such as grammar, vocabulary, and communicative competence are not only multidimensional but also dynamically realised during oral performance. Raters must attend to accuracy, range, fluency, coherence, and appropriateness while managing the flow of speech and, in interactive tasks, the speaker's responsiveness to interlocutors. This complexity increases the likelihood of selective attention to certain performance features at the expense of others, thereby introducing variability even when common scoring guidelines are provided [68].

Among the significant bias interactions, the marginal predominance of severity over leniency can be interpreted as a context-sensitive "standards-protection" response: when teachers perceive the assessment as consequential. They may default to stricter interpretations of descriptors to avoid grade inflation and to signal academic rigour. This stance is consistent with broader evidence that higher grading standards can be associated with improved downstream achievement, suggesting that some teachers may view stringent scoring as pedagogically productive rather than merely punitive [69]. However, the critical issue in speaking assessment is that "rigour" is not value-neutral, severity is often enacted through disproportionate attention to accuracy (grammar, lexical choice) and penalisation of errors that are salient to raters, even when communicative intent remains largely successful. In rater-mediated speaking contexts, such cue-weighting can systematically depress scores for students whose communicative competence is stronger than their formal accuracy, especially when the rubric partitions criteria but raters psychologically integrate them during judgment [70]. Evidence from speaking rating research shows that rater severity varies meaningfully as a function of rater characteristics and experience, reinforcing that harshness can emerge from rater cognition rather than from the performance alone [8].

At the same time, the presence of leniency, often linked to perceived effort, familiarity, or affective reactions, highlights the relational nature of classroom-based and teacher-mediated assessment. Teachers are not detached scorers. They frequently rate through a socio-emotional lens shaped by day-to-day interactions with learners, which can elevate the perceived "deservingness" of a performance even when it is linguistically weaker. This duality is intensified in speaking tasks because oral performance carries interpersonal signals, such as confidence and responsiveness, that can trigger positive global impressions and, in turn, spill over into analytic criteria. Empirically, halo processes remain a documented risk in analytic speaking scales: for instance, rated fluency can be influenced by non-fluency features such as lexical

complexity, demonstrating how a favourable (or unfavourable) overall impression can blur intended criterion boundaries [71]. In other words, leniency is not simply "kindness". It can be a systematic construct-irrelevant influence that reallocates score variance away from the targeted traits and toward interpersonal impressions.

The co-occurrence of severity and leniency supports a more critical conclusion, that rater behaviour is better understood as conditional rather than stable. Teachers oscillate between "guarding standards" and "supporting learners" depending on the cues they privilege, the task demands, and the interpretive latitude afforded by the rubric. This is why rater drift and rater effects persist even in ostensibly standardised systems; rater error is complex and difficult to estimate or control from limited scoring occasions, which complicates quality assurance if monitoring is infrequent or overly coarse [72]. Recent evidence also shows that different types of experience relate to rater severity and consistency, implying that "more experience" is not uniformly protective. What matters is which experience and how it calibrates internal standards. Practically, these findings justify design features that go beyond one-off training, such as ongoing calibration with benchmark performances, targeted feedback on severity/leniency tendencies, and analytic modelling of rater effects (e.g., MFRM) to detect bias patterns and adjust interpretations accordingly. They also support study designs that explicitly model rater–ratee interactions (not just main effects), because bias often manifests through relational cue combinations rather than through a single rater trait [63].

Mid-ability students accounted for a disproportionate share of significant bias interactions, a pattern that is theoretically plausible in rater-mediated speaking assessment because performances in the "middle bands" are more indeterminate and therefore place greater interpretive demands on raters. In analytic speaking rubrics, mid-range performances often show uneven profiles, for instance adequate vocabulary but unstable grammatical control, or acceptable accuracy but limited communicative effectiveness, which increases the likelihood that raters will rely on personal weighting of criteria, locally salient cues, and implicit thresholds when deciding between adjacent score bands [59]. This vulnerability is amplified by well-documented rater response tendencies such as centrality, the propensity to cluster ratings around the middle categories, which has been empirically examined in large-scale speaking performance assessment and is treated as a consequential rater effect for fairness monitoring [63]. Methodologically, evidence from multifaceted Rasch work also indicates that rater effects and bias patterns can vary systematically with examinee proficiency, meaning that the same rater may not behave consistently across ability levels [60]. From a validity and equity standpoint, this is consequential because mid-ability students are typically a substantial segment of the cohort. If rubric boundaries are not consistently applied and rater effects cluster at the "borderline" between adjacent categories, students with comparable speaking performance may receive less stable outcomes than peers whose performance is clearly low or clearly high [59, 63].

6 Conclusion

The findings advance theory, method, and practice in rater-mediated speaking assessment. Theoretically, they bolster a Lens Model view of rating as cue-mediated and interactional, where ambiguity in performance (especially among mid-ability students) catalyses construct-irrelevant variance that simple severity and leniency accounts cannot capture, thereby calling for refinements to cognitive and socio-constructivist models that explicitly incorporate student-driven bias mechanisms. Methodologically, the linked-design MFRM analysis demonstrates how calibrating raters, students, items, and domains can surface complex teacher x student biases that would likely remain hidden under CTT or broad reliability indices, and points toward mixed-methods extensions, such as think-alouds and eye-tracking, and facet-specific disaggregation across ability bands. Practically, the results argue for bias-aware professional development that goes beyond rubric familiarisation and embeds simulation-based calibration, exposure to diverse profiles, reflective scoring, and routine MFRM-based bias audits, alongside moderation and dual-rating procedures, and real-time "bias control" dashboards. This can monitor drift and improve assessment literacy, noting that a substantial share of ratings remained significantly skewed even after training.

Despite its valuable findings, this study has several limitations. The study was conducted within a specific context, focusing on secondary school English teachers in Malaysia, which limits the generalizability of the results to other educational systems or subject areas. Additionally, the study examined only a single type of assessment (English-language speaking assessment), and the observed rating biases may not extend to other forms of language or performance-based assessments. The study also relied on the MFRM model, which, while effective for identifying bias, may not capture all the factors that contribute to bias in teacher ratings. Finally, although rater training was conducted prior to the rating process, its effectiveness was not empirically examined in this study. Therefore, the extent to which the training enhanced rater consistency, reduced rater severity or leniency effects, and improved the usability of the scoring rubric remains a limitation that should be addressed in future studies.

Future research could pursue several avenues to deepen our understanding of rating bias. First, examining bias across different subjects would help determine whether the trends observed in language assessments are also present in other disciplines. Second, investigating bias across educational levels, such as primary and tertiary education, could reveal how rating bias changes as students age and as task complexity increases. Third, conducting longitudinal studies would shed light on the progression of teacher bias over time and assess the impact of professional development on rating practices. Fourth, exploring the influence of cultural and linguistic diversity on teacher bias would provide valuable insights into cross-cultural assessments. Finally, research should focus on evaluating the effectiveness of interventions and training programs to reduce rating bias and enhance fairness in assessments.

This study documents meaningful bias in teacher ratings of English speaking, predominantly severity effects, most evident in interactions involving lower-ability students and in the communication domain, the most challenging to judge. These findings reinforce the imperative for systematic, bias-focused professional learning and quality-assurance procedures so that ratings more faithfully reflect students' true abilities.

Acknowledgments. The authors gratefully acknowledge the Pacific Rim Objective Measurement Symposium (PROMS) 2025 Committee for recognizing this work with the Distinguished Student Scholar-ship Award, an honour that encouraged and amplified the study's scholarly impact. We also extend our sincere appreciation to Kumpulan Penilaian Pendidikan, Faculty of Education, Universiti Kebangsaan Malaysia (UKM), Bangi, for their invaluable guidance, constructive feedback, and steadfast institutional support throughout the conduct of this research.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. AERA, APA, NCME: Standards for Educational and Psychological Testing. American Educational Research Association, Washington (2014).
2. Guo, W., Wind, S.A.: The effects of rating designs on rater classification accuracy and rater measurement precision in large-scale mixed-format assessments. Appl. Psychol. Meas. 47, 91–105 (2023). https://doi.org/10.1177/01466216231151705.
3. Neittaanmäki, R., Lamprianou, I.: Communal factors in rater severity and consistency over time in high-stakes oral assessment. Lang. Test. 41, 584–605 (2024). https://doi.org/10.1177/02655322241239363.
4. Shahi, R., Ravand, H., Rohani, G.R.: Examining the effect of item difficulty and rater leniency on Iranian test takers' performance on WDCT and DSAT: A comparative study. Int. J. Lang. Test. 14, 1–19 (2024). https://doi.org/10.22034/ijlt.2024.454478.1341.
5. Mohd Noh, M.F., Mohd Matore, M.E.E.: 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. Front. Psychol. 13, 1–13 (2022). https://doi.org/10.3389/fpsyg.2022.941084.
6. Neittaanmäki, R., Lamprianou, I.: All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency. Lang. Test. 41, 606–626 (2024). https://doi.org/10.1177/02655322241239362.
7. Sandlund, E., Sundqvist, P.: Rating and reflecting: Displaying rater identities in collegial L2 english oral assessment. Assess. Speak. Context Expand. Constr. its Appl. 131–161 (2021). https://doi.org/10.21832/9781788923828-007.
8. Mohd Noh, M.F., Mohd Matore, M.E.E.: Rating performance among raters of different experience through multi-facet Rasch measurement (MFRM) model. J. Meas. Eval. Educ. Psychol. 11, 147–162 (2020). https://doi.org/10.21031/epod.662964.

9. Kang, O., Rubin, D., Kermad, A.: The effect of training and rater differences on oral proficiency assessment. Lang. Test. 36, 481–504 (2019). https://doi.org/10.1177/0265532219849522.
10. Engelhard, G., Wind, S.A.: Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments. Routledge, New York (2018). https://doi.org/10.1017/CBO9781107415324.004.
11. Brunswik, E.: The conceptual framework of psychology. Univ. Chicago Press, Chicago, IL (1952).
12. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1978).
13. Berger, P., Luckmann, T.: The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday & Company (1966).
14. Engelhard, G., Wang, J.: Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences. Routledge Taylor & Francis Group, New York & London (2024).
15. Blackhurst, T., Warmelink, L., Roestorf, A., Hartley, C.: The Brunswik Lens Model: a theoretical framework for advancing understanding of deceptive communication in autism. Front. Psychol. 15, 1–14 (2024). https://doi.org/10.3389/fpsyg.2024.1388726.
16. Bhowmik, C. V, Back, M.D., Nestler, S.: Appearing smart, confident and motivated: a lens model approach to judgment accuracy in an educational setting. Springer Netherlands (2025). https://doi.org/10.1007/s11218-025-10057-1.
17. Beck, J., Dutke, S., Utesch, T.: Understanding teacher judgments of student motivation: The role of (un-) available cues. Learn. Instr. 95, 1–11 (2025).
18. Asli, N.F., Mohd Matore, M.E.E.: Rater errors in rating process and the needs to be identified among student raters. J. Contemp. Issues Bus. Gov. 27, 6140–6158 (2021). https://doi.org/10.47750/cibg.2021.27.02.608.
19. Eckes, T.: Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. Peter Lang, Frankfurt (2015).
20. Erguvan, I.D., Dunya, B.A.: Analyzing rater severity in a freshman composition course using many facet Rasch measurement. Lang. Test. Asia. 10, 1–20 (2020). https://doi.org/10.1186/s40468-020-0098-3.
21. Chen, Q.: Rubricizing the assessment practice : A systematic review and meta-analysis of rubrics in rater-mediated assessment of language interpreting. Lang. Test. 1–3 (2025). https://doi.org/10.1177/02655322251391233.
22. Khamboonruang, A.: Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A Many-Facets Rasch Measurement approach. PASAA J. 66, 5–36 (2023). https://doi.org/10.58837/chula.pasaa.66.1.1.
23. Koizumi, R., In'nami, Y., Fukazawa, M.: Comparison between holistic and analytic rubrics of a paired oral test. JLTA J. 23, 57–77 (2020). https://doi.org/10.20622/jltajournal.23.0_57.
24. Limgomolvilas, S., Sukserm, P.: Examining rater reliability when using an analytical rubric for oral presentation assessments. Learn J. Lang. Educ. Acquis. Res. Netw. 18, 110–134 (2025).
25. Khabbazbashi, N., Galaczi, E.D.: A comparison of holistic, analytic, and part marking models in speaking assessment. Lang. Test. 37, 1–18 (2020). https://doi.org/10.1177/0265532219898635.
26. Mendoza, L.E.: Using rubrics for language assessment. In: Gonsalves, C. and Pearson, J. (eds.) Improving Learning Through Assessment Rubrics: Student Awareness of What and How They Learn. pp. 1–380. IGI Global Scientific Publishing (2023). https://doi.org/10.4018/978-1-6684-6086-3.

27. Aryadoust, V., Zakaria, A., Lim, M.H., Chen, C.: An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. Front. Psychol. 11, 1–29 (2020). https://doi.org/10.3389/fpsyg.2020.01941.
28. Ahmadi Shirazi, M.: For a greater good: Bias analysis in writing assessment. SAGE Open. 9, 1–14 (2019). https://doi.org/10.1177/2158244018822377.
29. Kwako, A., Wan, Y., Zhao, J., Chang, K.W., Cai, L., Hansen, M.: Does BERT exacerbate gender or L1 biases in automated English speaking assessment? In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 668–681 (2023). https://doi.org/10.18653/v1/2023.bea-1.54.
30. Sata, M., Karakaya, I.: Investigating the impact of rater training on rater errors in the process of assessing writing skill. Int. J. Assess. Tools Educ. 9, 492–514 (2022). https://doi.org/10.21449/ijate.877035.
31. Lyu, M.: A comparison of two scoring methods of Chinese university students' writing in EFL course using MultiFacet Rasch Model. In: Proceedings of the 2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022). pp. 298–304 (2022). https://doi.org/10.2991/assehr.k.220704.054.
32. Buyukkidik, S.: Purification procedures used for the detection of gender DIF: Item bias in a foreign language test. Int. J. Assess. Tools Educ. 10, 765–780 (2023). https://doi.org/10.21449/ijate.1250358.
33. Zhu, Y., Fung, A.S.L., Yang, L.: A methodologically improved study on raters' personality and rating severity in writing assessment. SAGE Open. 11, 1–16 (2021). https://doi.org/10.1177/21582440211009476.
34. Xie, X., Nimehchisalem, V., Yong, M.F., Yap, N.T.: Evaluating the accuracy of peer assessment of ESL argumentative writing using a mixed-methods approach. In: Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023). pp. 251–268. Atlantis Press International BV (2024). https://doi.org/10.2991/978-94-6463-494-5_16.
35. Yang, S., Wang, J.: Measuring scientific creativity with subjective creativity assessments: Psychometric challenges and suggestions. In: Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023). pp. 65–81. Atlantis Press International BV (2024). https://doi.org/10.2991/978-94-6463-494-5_4.
36. Doosti, M., Safa, M.A.: Fairness in oral language assessment: Training raters and considering examinees' expectations. Int. J. Lang. Test. 11, 64–90 (2021).
37. Aryadoust, V., Ng, L.Y., Sayama, H.: A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Lang. Test. 38, 6–40 (2021). https://doi.org/10.1177/0265532220927487.
38. Honko, M., Neittaanmäki, R., Jarvis, S., Huhta, A.: Beyond literacy and competency – The effects of raters' perceived uncertainty on assessment of writing. Assess. Writ. 57, 1–14 (2023). https://doi.org/10.1016/j.asw.2023.100768.
39. Polat, M., Akay, E.: Raters' prejudice in the assessment of oral performance. Eur. J. English Lang. Teach. 3, 1–14 (2017). https://doi.org/10.5281/zenodo.1067199.
40. Alp, P., Epner, A., Pajupuu, H.: The influence of rater empathy, age and experience on writing performance assessment. Linguist. Beyond Within. 3, 7–19 (2018).
41. Protivínský, T., Münich, D.: Gender bias in teachers' grading: What is in the grade. Stud. Educ. Eval. 59, 141–149 (2018). https://doi.org/10.1016/j.stueduc.2018.07.006.
42. Bijani, H., Khabiri, M.: Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment: A FACETS analysis. J. Asia TEFL. 14, 687–702 (2017). https://doi.org/10.18823/asiatefl.2017.14.4.7.687.
43. Liu, J., Xie, L.: Examining rater effects in a WDCT pragmatics test. Iran. J. Lang. Test. 4, 50–65 (2014).

44. 4Winke, P., Gass, S., Myford, C.: Raters' L2 background as a potential source of bias in rating oral performance. Lang. Test. 30, 231–252 (2013). https://doi.org/10.1177/0265532212456968.
45. Yan, X.: An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. Lang. Test. 31, 501–527 (2014). https://doi.org/10.1177/0265532214536171.
46. Isbell, D.R.: Assessing C2 writing ability on the certificate of english language proficiency: Rater and examinee age effects. Assess. Writ. 34, 37–49 (2017). https://doi.org/10.1016/j.asw.2017.08.004.
47. Kim, Y.H.: An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. Lang. Test. 26, 187–217 (2009). https://doi.org/10.1177/0265532208101010.
48. Jeong, H.: Narrative and expository genre effects on students, raters, and performance criteria. Assess. Writ. 31, 113–125 (2017). https://doi.org/10.1016/j.asw.2016.08.006.
49. Razak, R.A., Yusoff, S.M., Leng, C.H., Marzaini, A.F.M.: Evaluating teachers' pedagogical content knowledge in implementing classroom-based assessment: A case study among ESL secondary school teachers in Selangor, Malaysia. PLoS One. 18, 1–23 (2023). https://doi.org/10.1371/journal.pone.0293325.
50. Halim, H.A., Hamzah, M.I., Zulkifli, H.: A systematic review on the formative assessment practice in teaching and learning in secondary school. Int. J. Eval. Res. Educ. . 13, 1173–1183 (2024). https://doi.org/10.11591/ijere.v13i2.26187.
51. Rahim, S.S.B.A., Chua, K.H.: Basic education in Malaysia. In: Symaco, L.P. and Hayden, M. (eds.) International Handbook on Education in South East Asia. pp. 591–607. Springer Nature Singapore, Singapore (2024). https://doi.org/10.1007/978-981-16-8136-3_11-1.
52. Dreer-Goethe, B.: Well-being and mentoring in pre-service teacher education: An integrative literature review. Int. J. Mentor. Coach. Educ. 12, 336–349 (2023). https://doi.org/10.1108/IJMCE-09-2022-0073.
53. Skerritt, C.: School autonomy and the surveillance of teachers. Int. J. Leadersh. Educ. 26, 553–580 (2023). https://doi.org/10.1080/13603124.2020.1823486.
54. Mohamad Marzaini, A.F., Haslee Sharil, W.N.E., Supramaniam, K., Mat Yusoff, S.: Evaluating Teachers' assessment literacy in enacting CEFR-aligned classroom-based assessment in Malaysian secondary schools ESL classroom. Int. J. Acad. Res. Progress. Educ. Dev. 12, 611–624 (2023). https://doi.org/10.6007/ijarped/v12-i1/15691.
55. Anthony, L., Miriam, S.: Praktis Intensif: English Language Form 3 KSSM. EPH Publishing, Kuala Lumpur (2020).
56. Lawshe, C.H.: A quantitative approach to content validity. Pers. Psychol. 28, 563–575 (1975).
57. Schaefer, E.: Rater bias patterns in an EFL writing assessment. Lang. Test. 25, 465–493 (2008). https://doi.org/10.1177/0265532208094273.
58. Wang, J., Luo, K.: Evaluating rater judgments on ETIC Advanced writing tasks : An application of generalizability theory and Many-Facets Rasch Model. Pap. Lang. Test. Assess. 8, 91–116 (2019).
59. Dimova, S.: Performance-based speaking tests: Possibilities in local language testing. Lang. Teach. Res. Q. 29, 120–133 (2022). https://doi.org/10.32038/ltrq.2022.29.08.
60. Bijani, H., Hashempour, B., Ahmed, K., Al, A., Said, S., Orabah, B., Heydarnejad, T.: Investigating the effect of classroom - based feedback on speaking assessment: A multifaceted Rasch analysis. Lang. Test. Asia. (2022). https://doi.org/10.1186/s40468-022-00176-3.

61. Bond, T.G., Yan, Z., Heene, M.: Applying the Rasch Model. Routledge, New York & London (2021). https://doi.org/10.4324/9781410614575.
62. Boone, W.J., Staver, J.R.: Advances in Rasch Analyses in the Human Sciences. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43420-5_21.
63. Jin, K., Eckes, T.: Detecting rater centrality effects in performance assessments: A model-based comparison of centrality indices. Meas. Interdiscip. Res. Perspect. 20, 228–247 (2022). https://doi.org/10.1080/15366367.2021.1972654.
64. Barkaoui, K.: Effects of marking method and rater experience on ESL essay scores and rater performance. Assess. Educ. Princ. Policy Pract. 18, 279–293 (2011). https://doi.org/10.1080/0969594X.2010.526585.
65. Myford, C., Wolfe, E.W.: Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part Il. J. Appl. Meas. 5, 189–227 (2004).
66. Wind, S.A.: Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? Assess. Writ. 43, 100416 (2020). https://doi.org/10.1016/j.asw.2019.100416.
67. Jong, N.H. De: Assessing second language speaking proficiency. Annu. Rev. Linguist. 9, 541–560 (2023). https://doi.org/10.1146/annurev-linguistics-030521- 052114.
68. Sasayama, S., Norris, J.M.: Designing speaking tasks for different assessment goals The complex relationship between cognitive task complexity, language performance, and task accomplishment. TASK. 2, 184–217 (2022). https://doi.org/10.1075/task.21020.sas.
69. Gershenson, S., Holt, S.B., Tyner, A.: Making the grade: The effect of teacher grading standards on student outcomes. Contemp. Econ. Policy. 42, 305–318 (2024). https://doi.org/10.1111/coep.12637.
70. Yu, K., Thomas, J.: When raters generalize : Examining sources of halo effects with mixture Rasch facets models. Behav. Res. Methods. 57, 1–19 (2025). https://doi.org/10.3758/s13428-025-02667-6.
71. Grady, S.O.: Halo effects in rating data: Assessing speech fluency. Res. Methods Appl. Linguist. 2, 1–12 (2023). https://doi.org/10.1016/j.rmal.2023.100048.
72. White, M., Ronfeldt, M.: Monitoring rater quality in observational systems : Issues due to unreliable estimates of rater quality. Educ. Assess. 29, 124–146 (2024). https://doi.org/10.1080/10627197.2024.2354311.

Appendix A: Speaking Test Item
Adapted from [56]

Item 1: Background Interview

- Good morning / afternoon / evening. My name is [interlocutor's name].
- First, I would like to ask you some questions about yourself.
- What is your name?
- What should I call you?
- Thank you.
- Where do you live?
- Is your house far from here?
- Do you enjoy studying English?
- How often do you have English lessons?
- When do you usually do your homework?
- Do you usually do your homework at school or at home?
- Tell me about your favourite song.
- Where do you usually go with your friends?
- Thank you. That is the end of Item 1.

Item 2: Storytelling

In this part of the test, I'm going to give each of you five pictures which tell a story. I'd like you to talk about them on your own for about a minute.

Candidate A, it's your turn first. Here are your pictures. You now have some time to look at the pictures.

Okay, are you ready (Candidate A)? Now, tell us the story in the pictures. Thank you. Can I have the booklet please? (Retrieve Candidate booklet).
Now, Candidate B, here are your pictures. You now have some time to look at the pictures.

Okay, are you ready (Candidate B)? Now, tell us the story in the pictures. Thank you. Can I have the booklet please? (Retrieve Candidate booklet).

Pictures for Candidate A



Pictures for Candidate B



Item 3: Discussion

Now, I'd like you to talk together for about two minutes. I am going to describe a situation to you. Your school is organising a jungle trekking activity for students.

(Place Candidate booklet, open at Item 2, in front of the candidates.)

Here are some things that students should bring when going jungle trekking. First, you have some time to look at the task.

(Allow candidates 15 seconds to look at the task.)

Now, talk together about the different things students should bring when going jungle trekking and decide which one would be the most important.

Okay, are you ready? Please start. Candidates

.....

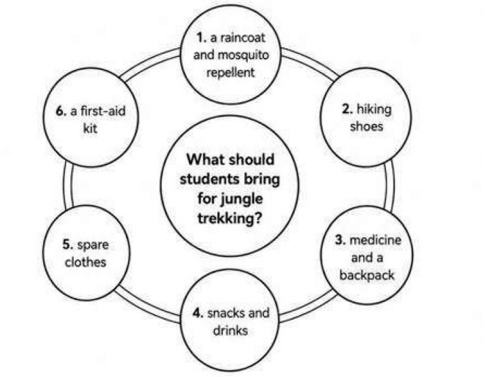
Back-up prompts to be used only if necessary:

- What do you think, (candidate name)?
- How about you, (candidate name)?
- What about this? (pointing to option)
- Do you agree, (candidate name)?

Thank you. That's the end of the test.

Can I have the booklet, please? *(Retrieve Candidate booklet.)*

The Mind Map



Appendix B: Scoring Rubric

Scores	Grammar	Vocabulary	Communicative Competence
5	• Can show a good degree of control of simple grammatical structures. • Can form longer sentences and link them together using basic cohesive devices.	• Can use a range of appropriate vocabulary to talk about everyday situations and familiar topics. • Can convey relevant meaning but may have errors in vocabulary choice.	• Can express him/herself clearly though there is some hesitation. • Can initiate, maintain, develop and close a conversation with little difficulty. • Can ask for clarification and further details in order to move a discussion forward.
4	Candidate performance has aspects of Band 3 and Band 5 performance.		
3	• Can show sufficient control of simple grammatical structures. • Can use connectors to link simple sentences.	• Can use sufficient vocabulary to talk about everyday situations and familiar topics. • Can mostly convey meaning but may use inappropriate vocabulary or some repetition.	• Can make him/herself understood but may have pauses and false starts. • Can initiate, maintain and close a conversation with some difficulty. • Can ask for clarification when necessary.
2	Candidate performance has aspects of Band 1 and Band 3 performance.		
1	• Can show only limited control of simple grammatical structures. • Can link words or phrases with very basic connectors.	• Can use basic vocabulary to talk about everyday situations and familiar topics. • Can produce responses with frequent repetition of words and phrases.	• Can manage short utterances with much pausing. • Can initiate, maintain and close a conversation with difficulty and using rehearsed phrases. • Can indicate with words, intonation and gestures that he/she does not understand.
0	Candidate does not produce sufficient language to be assessed. Candidate uses language(s) other than English.		

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

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

