Evaluating the Accuracy of Peer Assessment of ESL Argumentative Writing Using a Mixed-Methods Approach

Xiao Xie[1[0000-0003-3388-7464]] Vahid Nimehchisalem[2[0000-0002-5454-1895]]

Mei Fung Yong[3[0000-0002-5363-1864]] and Ngee Thai Yap[4[0000-0001-8672-7128]]

[1,2,3,4] Universiti Putra Malaysia, Persiaran Universiti 1 Serdang, Malaysia
vahid@upm.edu.my

Abstract. In the context of ESL writing, the prevalent approach of utilising inter-rater reliability measures, particularly Pearson's *r* coefficient, for the scrutiny of peer assessment comes with inherent constraints. Rasch models have emerged as an alternative method to conventional correlation analysis for assessing rater accuracy, as they show the absolute match between peer ratings and expert ratings, and compute individual-level statistics for each element of each assessment facet. This study aims to evaluate the accuracy of peer assessment regarding ESL argumentative writing, and to explore why some writing domains are difficult for peer raters to score accurately. Peer assessment training was conducted over a five-week period with 24 undergraduate students enrolled in an ESL argumentative writing course at a Malaysian university. A mixed-methods approach was used to examine the relationship between peer raters' quantitative ratings and their judgemental process. The quantitative data were analysed using Rasch Partial Credit Model (PCM), and the qualitative data were examined using constant comparative method and thematic analysis. The quantitative analyses reveal that the domain of Relevance and Adequacy of Content (RAC) was most likely to peer assess accurately, while the other two domains, Compositional Organisation (CO) and Cohesion (C) were most difficult to assess accurately by this cohort of peer raters. The qualitative analyses suggest that peer raters' justifications for their scores were not consistent with the reasoning of expert raters, partially explaining inaccurate ratings in certain domains. This comprehensive information has the potential to improve peer assessment in an accurate and consistent manner, and to better organise peer assessment in tertiary ESL writing training programmes.

Keywords: Mixed-Methods Approach, Partial Credit Model, Peer Assessment, Rater Accuracy, Rater Perception.

1 Introduction

Peer assessment is defined as 'an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*,
Atlantis Highlights in Social Sciences, Education and Humanities 23,
https://doi.org/10.2991/978-94-6463-494-5_16

of peers of similar status' (Topping, 1998, p. 250). Its role extends to acquainting students with the diverse expectations of quality held by distinct user groups, thus offering them a comprehensive grasp of the benchmarks and criteria applied to gauge their own work. This immersive involvement in evaluating the work of their peers not only cultivates their awareness of quality parameters but also exposes them to a gamut of examples and varying levels of accomplishment (Han, 2018). Within the domain of higher education, the exploration of peer assessment accuracy has commanded significant scholarly attention, driven by several pivotal factors (Han & Zhao, 2021). Firstly, the aptitude of students to accurately and consistently evaluate their peers bears testament to the successful internalisation and application of the instructor's grading criteria. Secondly, peer assessment serves as an avenue for students to gain deeper insights into and adeptness with scoring rubrics, fostering their comprehension and confidence in their use. Lastly, instances of erroneous peer assessments accentuate the necessity to scrutinise elements such as suboptimal peer assessment methodologies and latent biases among students.

Despite the potential benefits of peer assessment, researchers remain concerned about the reliability of peer ratings (Falchikov & Goldfinch, 2000; Li et al., 2016). It is worth noting that the prevalent approach of utilising inter-rater reliability measures, particularly Pearson's *r* coefficient, for the scrutiny of rater behaviour comes with inherent constraints (Han, 2018; Han & Zhao, 2021). One pivotal concern lies in the fact that the inter-rater reliability coefficient between two raters, A and B, lacks specificity towards each individual rater. Additionally, the application of Pearson's *r* merely indicates the likeness in scoring performances between raters, disregarding any potential statistically significant disparities that might persist. To surmount these limitations, the proposition of Rasch-based models has emerged as an alternative to the conventional correlation analysis. Yet, it is imperative to acknowledge that the implementation and suitability of these models in assessing the precision of peer ratings of ESL argumentative writing warrant further exploration.

This study aims to evaluate rater accuracy in ESL argumentative writing peer assessment, and to explore why some domains were difficult for peer raters to score accurately. By adopting a mixed-methods design (Wang et al., 2017), the quantitative component of the present study compares the challenges of peer assessing diverse domains accurately, while the qualitative component of the study investigates peer raters' justifications of scoring decisions when their ratings are not congruent with those provided by expert raters. The following research questions were addressed:

1. Which analytical assessment domains demonstrate higher likelihood of being peer assessed accurately?

2. What is the difference between expert raters' and peer raters' justifications of scoring decisions with regard to each domain of argumentative writing?

2 Literature Review

2.1 MFRM analysis of peer assessment of ESL writing

Four empirical studies examined peer assessment using the Many-Facet Rasch Measurement (MFRM), to evaluate the direct and indirect indicators of rater agreement, rater error and bias (Esfandiari & Myford, 2013; Farrokhi et al., 2012; Matsuno, 2009; Saito & Fujita, 2004). Saito and Fujita (2004) studied 47 students taking an English writing course at a Japanese university. The outcome of this comprehensive evaluation revealed a robust correlation between the ratings assigned by peers and those granted by teachers, resulting in a correlation coefficient of 0.72. There was, however, a nuanced discrepancy: teachers tended to adopt a more stringent grading approach than peer raters. In Matsuno (2009), 91 Japanese university students and four teachers were studied. Among students, a recurrent tendency emerged whereby they overestimated their peers' abilities while simultaneously underestimating their own. Contrary to this, teachers' assessments showed a more balanced disposition. In peer assessment, spelling was most generously evaluated, suggesting a degree of leniency; in contrast, grammar was most critically evaluated, indicating an increased level of scrutiny. In part, this could be due to the Japanese educational system, which places a great deal of emphasis on English grammar. A study by Farrokhi et al. (2012) examined 188 Iranian English majors at two state universities. As seen in the results, teachers tended to take a more severe approach to their assessments, especially when evaluating parameters such as reframing the topic, developing the thesis, and writing the introduction sections. Teachers demonstrated a contrasting tendency toward leniency in evaluating logic sequence and vocabulary usage when compared with students. Esfandiari and Myford (2013) analysed 194 participants from two Iranian national universities and drew the conclusion that cultural reluctance to critique classmates may be at the root of peer assessment's tendency to overestimate rather than underestimate, because of cultural beliefs, social dynamics, and religious principles.

The MFRM approach adopted by the four studies above sheds light on how peer assessment and teacher assessment align. However, it's crucial to clarify that this alignment does not necessarily indicate absolute concordance among the assessors. Instead, it offers an indirect gauge of the precision of peer assessment. While the application of Rasch modelling of rater accuracy emerges as a promising avenue, capable of providing a direct and granular examination of the accuracy inherent in peer assessment. In essence, the pursuit of a direct Rasch modelling approach becomes an imperative in gauging the true accuracy of peer assessment.

2.2 Rasch Modelling of Rater Accuracy

Rater accuracy is defined as 'the match between ratings obtained from operational raters and those obtained from an expert panel on a set of benchmark performance' (Engelhard, 1996, p. 57), regarding 'absolute match between operational ratings and expert ratings' (Han, 2018, p. 12). Brunswik's (1952) Lens Model is deeply rooted within a probabilistic functionalist framework and offers a useful conceptual framework for understanding the rater-mediated assessment accuracy. An ecosystem, a judgement system, and multiple cues are used in the Lens Model to demonstrate how various factors influence the decisions of raters, including the characteristics of the raters, the characteristics of tasks, and the features of student performance.

Specifically, Rasch modelling of rater accuracy includes key models such as the Dichotomous Model, Rating Scale Model (RSM), and Partial Credit Model (PCM) (Aryadoust et al., 2019; Han, 2018). The models allow a systematic analysis of how raters' responses align with the underlying constructs being measured. Rasch RSM assumes a fixed distance between scale category thresholds. Thus, thresholds remain uniform across different rating domains. Rasch PCM, on the other hand, assumes that each rating domain has its own unique scale structure, allowing each rating domain to have its own threshold parameterisation, which allows for greater flexibility in the approach. Intricacies of the assessment context determine whether RSM or PCM are appropriate. Rasch PCM often takes precedence when dealing with multiple rating domains when investigating rater accuracy. It is clear that this is the case from the study of Han (2018), as well as the study of Han and Zhao (2021), where the Rasch PCM approach was used to examine rater accuracy in the context of translation evaluation with a variety of rating domains, acknowledging the fact that Rasch PCM is also widely applied in the evaluation of writing and speaking.

3 Methodology

3.1 Participants

We recruited 24 undergraduate students from a Malaysian public university, mainly females aged between 19 and 21, using a convenience sampling method. They had achieved at least Band 3 on the Malaysian University English Test (MUET), demonstrating a reasonably fluent and fairly appropriate use of English language, despite numerous grammatical errors. Most of their bands corresponded to CEFR levels B1 and B2. It can be concluded that these participants possessed intermediate English proficiency, enabling them to write clear, detailed texts about a variety of topics. Additionally, they could articulate viewpoints on topical issues, presenting advantages and disadvantages.

3.2 Instrument

The seven-point, descriptor-based analytical rating scale (Aryadoust, 2012) was adopted to evaluate five domains of writing: Relevance and Adequacy of Content (RAC), Compositional Organisation (CO), Cohesion (C), Vocabulary (V), and Grammar (G). Aryadoust (2012) opted for a broader scale with seven levels of language proficiency, as opposed to the narrower TEEP scale, to capture a wider range of proficiency levels. In accordance with Weir's (1990), RAC measures sociolinguistic knowledge, assessing the relevance of the response to the topic of interaction or task setting; V and G encompass linguistic knowledge, assessing the range and accuracy of lexical and grammatical knowledge; and CO and C reflect discourse knowledge, assessing the organisation and flow of arguments as well as the cohesion and coherence of ideas in the essays. Aryadoust (2012) conducted a MFRM analysis, and the statistical evidence confirmed the overall utility of the scale.

3.3 Training Procedure

Peer assessment training lasted from week 2 to week 6, encompassing five weeks, throughout the second semester of the academic year 2022/2023. There was a total of 15 hours of training provided to the participants during the three-round peer assessment programme, which consisted of two 1.5-hour sessions per week. At the beginning of the training, the first author presented various exemplars to illustrate the assessment standards for each domain of the analytical rating scale by Aryadoust (2012). Subsequently, before the next training session, participants were asked to write one of the three writing tasks in 300-500 words: (1) Music's role in bringing people of different cultures and age groups together; (2) The potential obsolescence of print newspapers and books due to online reading; (3) The challenges of living in a foreign-language nation. After accomplishing each writing task, participants were guided with in-class discussion of two writing exemplars representing different proficiency levels in the following session. The exemplars' diverse writing performances were analysed by the first author, aiming to further cultivate participants' evaluative judgement. After fulfilling the aforementioned scaffolding procedures, participants would complete the in-class peer assessment and write the rating reflections in the subsequent session. After collecting the quantitative and qualitative data, the next round of peer assessment continued.

3.4 Expert Rating

In this study, expert raters were three university lecturers who had earned postgraduate degrees in English Language Education and had taught English academic writing at higher education institutions for an average of eight years. Their experience also included formative and summative assessments of writing performance.

A brief training session was conducted to familiarise the expert raters with the analytical rating scale by Aryadoust (2012). The first author emailed each rater with the rating scale and asked them to familiarise themselves with the criteria. In addition, an online training session explained scoring forms and rating criteria, in which the first author demonstrated how to assess argumentative writing samples simultaneously. Any misunderstandings or ambiguities were clarified during this demonstration to ensure consistency. As a result of the training process, the expert raters' understanding and application of the scoring criteria were aligned. It enhanced the reliability and validity of our findings by ensuring a standardised and objective evaluation of selected essays.

After the training session, the expert raters were asked to assign scores to 18 writing essays of the three tasks. Cronbach's α (0.90) indicates a high level of agreement among the three expert raters. Then, the first author averaged the three expert raters' scores and rounded them to whole numbers to estimate true scores for each domain of the essay.

To address any inconsistencies and gather additional insight into their scoring results, the expert raters were requested to provide additional reflections on the five domains to justify their scoring decisions, shedding light on the reasoning behind their ratings and adding depth and richness to the data analysis process.

3.5 Data Collection

For each round of peer assessment, all participants were instructed to evaluate the six essays with varying writing proficiency levels, based on the five domains of the analytical rating scale by Aryadoust (2012). Peer assessment in a structured classroom setting allowed for thoughtful evaluation. As a result, participants could concentrate solely on the assessment process without distractions that could compromise their integrity and accuracy.

In class, participants were asked to reflect on their scoring decisions regarding each domain of the selected essays after completing each round of peer assessment. The purpose of these rating reflections is to gain deeper insights into the perceptions and justifications behind the scoring decisions made by the participants.

3.6 Data Analysis

The data analysis procedure for Research Question 1 followed three steps (Han, 2018; Han & Zhao, 2021). Firstly, the raw scores of peer assessment were transformed into rater accuracy indices, then Rasch PCM was utilised to calculate the rater accuracy with the FACETS programme version 3.80.0, and illustrative data analyses were conducted with reference to Engelhard (1996, 2013), Engelhard and Wind (2017), as well as Wind and Engelhard (2013). It was deemed appropriate to align peer assessment results with expert rating, given that the scoring criteria were clearly defined and understood by both peer raters and expert raters. The peer assessment rater accuracy index (X_{ni}) was computed using the following formula (see Engelhard, 1996):

$$x_{ni} = \max \{ |S_{ni} - I_i| \} - |S_{ni} - I_i| \tag{1}$$

Here, S_{ni} represents the observed score of peer rater n for writing item i, I_i denotes the score given by the expert rating panel for writing item i, and max {...} signifies the maximum possible value for n and i. The seven categories ranging from 1 to 7 of the analytical rating scale by Aryadoust (2012) were utilised, which resulted in 13 possible differences between expert scores and observed peer assessment scores (i.e., -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6), and seven absolute values ranging from 0 to 6. It is crucial to note that the rater accuracy index (X_{ni}) increases with greater levels of accuracy in the peer assessment.

Moving on to the second step, the Rasch PCM was utilised, referring to the following mathematical formula, which includes four facets: the rater facet, as a function of peer raters' ability to assign accurate scores; the writing task facet, the difficulty of accurately scoring specific writing tasks; the essay facet, the difficulty in accurately scoring specific essays; the domain facet, the difficulty in accurately scoring specific rating domains. Through a joint maximum likelihood estimation procedure, the Rasch PCM generated calibrated estimates for all elements of each scoring facet.

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = \beta_n - \lambda_i - \delta_j - \nu_m - \gamma_k \tag{2}$$

where

P_{nijmk} represents the probability of peer rater n accurately scoring the domain m of essay j of writing task i as category k.

$P_{nijmk-1}$ represents the probability of peer rater n accurately scoring the domain m of essay j of writing task i as category $k-1$.

β_n represents the scoring accuracy of peer rater n.

λ_i represents the difficulty of accurately scoring writing task i.

δ_j represents the difficulty of accurately scoring essay j.

ν_m represents the difficulty of accurately scoring writing domain m.

γ_k represents the difficulty of accurately scoring category k relative to category $k-1$.

Lastly, the first author examined three main categories of statistical indices (Han, 2018; Han & Zhao, 2021), including logit-scale location (i.e., the Rasch-calibrated accuracy measures in logits, and the standard error), separation (i.e., the chi-square statistic, and the reliability of separation), and data-model fit (i.e., Infit MnSq, and Outfit MnSq).

To address Research Question 2, this study utilised a thematic analysis approach, which involved a harmonious blend of inductive and deductive coding methods, aimed at gaining a comprehensive understanding of the identified theme (Fereday & Muir-Cochrane, 2006). With the help of NVivo software (1.5.1), the qualitative data for this study was analysed by examining expert rating reflections as well as peer assessment reflections. Furthermore, the constant comparative method (Wang et al., 2017) was employed to conduct a re-analysis of the reflection responses. This re-analysis involved matching the responses to the rating domain and determining the frequency of each response within the domain. The aim was to investigate whether specific essay characteristics were associated with the perceptions of expert raters and peer raters, shedding light on why certain essays posed challenges for accurate rating. To gauge the accuracy of ratings, the essays were ranked based on the number of ratings that aligned with the criterion scores provided by three expert raters. A lower count of accurate ratings indicated a greater difficulty in accurately rating a particular domain of the essay. The study distinguished between accurate ratings, which matched the standard ratings, and two types of inaccurate ratings: above-standard ratings (ratings higher than the criterion scores) and below-standard ratings (ratings lower than the criterion scores). By examining the reasons provided in the reflection responses, the first author sought to identify the commonly cited characteristics of essays as identified by both peer raters and experts. This systematic analysis provided insights into the factors influencing accurate rating and highlighted the most frequently mentioned characteristics identified by both groups.

4 Results

4.1 Quantitative Data Analysis

Table 1 presents a summary of the statistics pertaining to Rasch PCM, comprising four facets: writing task, writing essay, peer rater, and rating domain. To determine whether

these facets are statistically significant, the chi-square (χ^2) indices and reliability of separation (Rel) values should be considered. In relation to the writing task facet, the chi-square index (χ^2=4.6, $p > 0.05$) and the reliability of separation (Rel=0.35) reveal that the difficulties for three writing tasks to peer assess accurately do not have statistically significant difference. Conversely, in relation to the other facets, which focus on writing essays, peer raters, and rating domains, the chi-square (χ^2) indices reveal statistically significant differences. Additionally, the values of reliability of separation (Rel) reveal that each element within these following facets could be reliably distinguished.

As for the rating domain facet, the statistical analysis, as indicated by the chi-square index ($\chi^2 = 67.5$) in Table 1, confirms a significant difference in rating accuracy across the various rating domains. This finding underscores the importance of recognising the varying levels of difficulty that peer raters experience when evaluating different domains of argumentative essays. The reliability of separation (Rel=0.92), further reinforces this finding that each element within the rating domain facet can be reliably distinguished from one another.

Table 1. Summary statistics for peer assessment accuracy model.

	Assessment Facets			
Logit Scale	Writing Task	Writing Essay	Peer Rater	Rating Domain
M	0.06	0.50	0.39	0.23
SD	0.04	0.48	0.37	0.22
N	3	18	24	5
Chi-Square (χ^2)	4.6	290.7*	173.1*	67.5*
Degree of Freedom	2	17	23	4
Reliability of Separation (Rel)	0.35	0.94	0.87	0.92

Note: Each asterisk (*) indicates $p < 0.05$

Figure 1 presents a Wright Map, providing a graphical representation of the calibrated estimates in logit for all assessment facets: writing task, writing essay, peer rater, and rating domain. The first column of the graph represents the true interval logit scale, mapping all calibrated facets and associated elements. For the second column, one can observe the logit estimates of the overall difficulty for each writing task to peer assess accurately. The third column provides logit estimates of the overall difficulty for each of the 18 writing essays to peer assess accurately, spanning approximately 2 logits (ranging from -1.0 logit to 1.0 logit). This wide range of variation indicates significant differences in the levels of accuracy demonstrated by the raters across the various writing essays. The fourth column provides a concise summary of the calibrated estimates for the 24 peer raters. Each asterisk (*) within the column symbolises an individual peer rater, arranged in descending order of accuracy. Remarkably, the broad distribution of rater accuracy estimates suggests that there is a notable diversity among this cohort of peer raters. Some raters consistently demonstrate higher levels of accuracy, positioning

them towards the top of the column, while others exhibit lower levels of accuracy, appearing towards the bottom. The fifth column demonstrates the calibrated estimates for the overall accuracy for the five domains, with the most accurate domain on the top and the least accurate at the bottom. Notably, the RAC domain emerges as the most likely to peer assess accurately, while the domains of C and CO pose greater challenges for peer raters to assess accurately. Lastly, the sixth to tenth columns demonstrate the scale of rating accuracy for the five domains, showing how these domains did not share a common structure, so the use of Rasch PCM in analysing the data is appropriate.

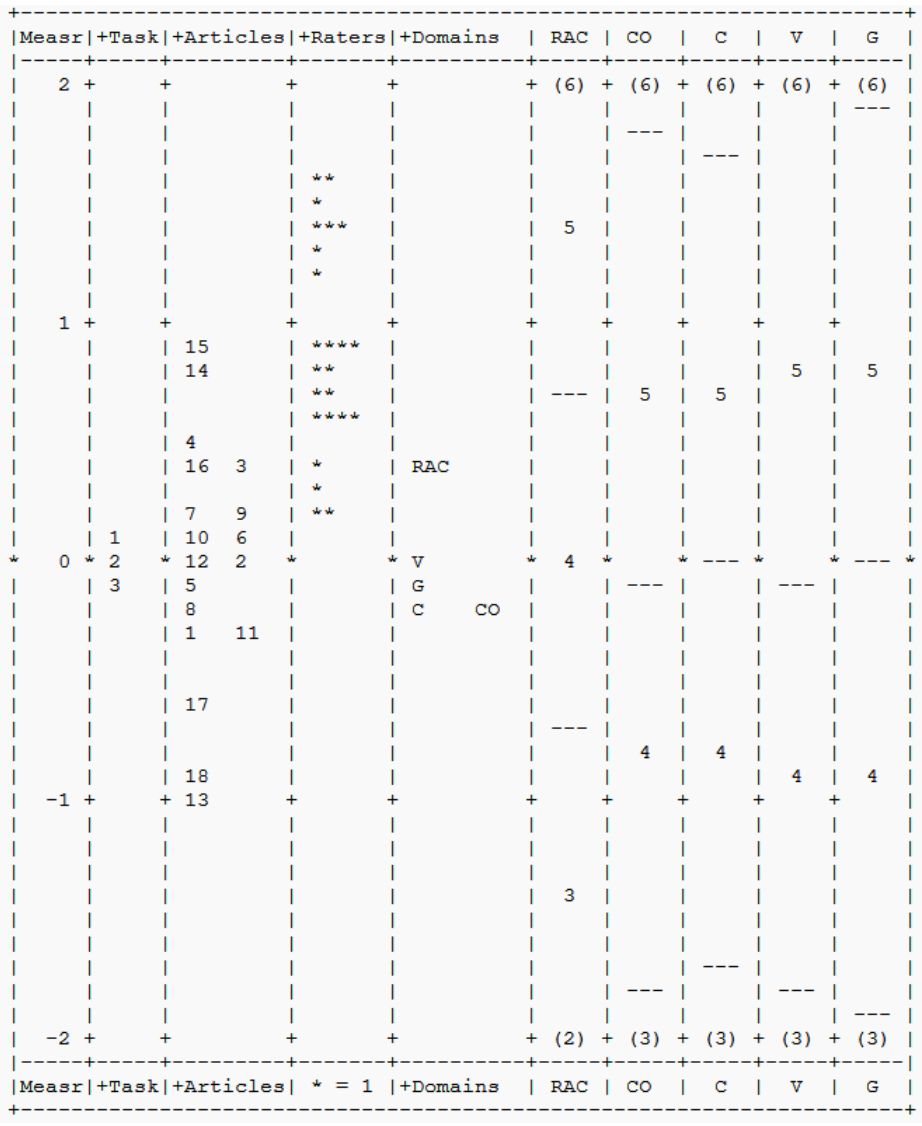

Fig. 1. Wright map of peer assessment accuracy model.

As shown in Table 2, in accordance with the Wright map, RAC domain obtains a logit estimate of 0.44, indicating that it is the most likely to be scored accurately by peer raters. On the other hand, the CO and C domains obtain logit estimates of -0.20 and -0.19 respectively, signifying that they are the most difficult to peer assess accurately.

Table 2. Calibration of the writing domain facet.

Rating Domain	Rating Accuracy Estimate (in logit)	Model S.E.	Infit MnSq	Outfit MnSq
RAC	0.44	0.06	1.06	1.08
V	0.04	0.07	0.97	0.97
G	-0.09	0.07	0.97	0.98
C	-0.19	0.06	1.05	1.04
CO	-0.20	0.07	0.89	0.89

Note: S.E. = standard error; MnSq = mean square.

4.2 Qualitative Data Analysis

Table 3 shows ratings for Essays on the RAC domain. It can be concluded that essays with higher criterion scores are more likely to be accurately rated by a larger number of peer raters. It is also found peer raters frequently overrate rather than underrate when analysing inaccurate ratings for then RAC domain.

Table 3. Ratings on essays for RAC domain (ordered from high to low accuracy).

Essay No.	Criterion Scores	Accurate Ratings	Inaccurate Ratings	
			Above	Below
14	6	13	2	9
16	6	13	7	4
15	6	12	7	5
04	6	9	12	3
09	4	9	14	1
10	5	8	9	7
06	4	7	10	7
02	5	6	18	0
05	3	5	19	0
07	5	5	16	3
08	5	5	9	10
03	4	4	17	3
18	4	4	20	0
01	5	3	19	2

11	4	3	19	2
17	4	3	21	0
12	3	2	22	0
13	3	1	22	1

Table 4 shows the ratings for essays on the CO domain. Similar to the RAC domain, essays with higher criterion scores receive more accurate ratings from a larger number of peer raters in the CO domain. Observing peer raters' inaccurate score distributions, it is also found that they often overrated rather than underrated for the CO domain.

Table 4. Ratings on essays for CO domain (ordered from high to low accuracy).

Essay No.	Criterion Scores	Accurate Ratings	Inaccurate Ratings	
			Above	Below
15	6	12	5	7
14	6	11	4	9
16	6	10	9	5
03	5	10	11	3
10	5	8	11	5
12	4	8	13	3
02	6	7	13	4
06	4	7	7	10
08	5	7	14	3
05	4	6	15	3
09	4	6	16	2
01	5	5	17	2
04	6	4	17	3
07	5	4	18	2
11	4	3	20	1
13	4	2	22	0
17	4	1	23	0
18	4	0	24	0

Table 5 shows the ratings for essays on the C domain. Unlike the RAC and CO domains, essays with higher criterion scores in this domain do not possess a distinctive tendency to receive more accurate ratings from peer raters. It is also found that peer raters often overrated rather than underrated for the C domain, based on their inaccurate score distributions.

Table 5. Ratings on essays for C domain (ordered from high to low accuracy).

Essay No.	Criterion Scores	Accurate Ratings	Inaccurate Ratings

			Above	Below
15	6	13	6	5
03	5	12	10	2
05	4	12	9	3
12	4	10	11	3
07	5	9	13	2
09	4	8	16	0
02	6	7	16	1
14	5	7	10	7
17	4	7	17	0
06	4	6	7	11
10	6	6	4	14
11	4	6	16	2
01	5	5	17	2
13	5	5	14	5
16	5	5	19	0
04	6	3	15	6
08	4	1	22	1
18	4	1	23	0

Table 6 shows essay ratings for the V domain. Unlike the RAC and CO domains, essays with higher criterion scores in this domain do not tend to receive more accurate ratings from peers. Based on their inaccurate score distributions, a substantial portion of these scores were assigned above the criterion score, indicating that peer raters also overrated rather than underrated for the V domain.

Table 6. Ratings on essays for V domain (ordered from high to low accuracy).

Essay No.	Criterion Scores	Accurate Ratings	Inaccurate Ratings	
			Above	Below
14	5	11	6	7
16	6	11	8	5
18	5	10	14	0
03	5	9	10	5
05	4	9	12	3
12	5	9	4	11
10	6	8	2	14
02	5	7	12	5
09	4	7	17	0
15	5	7	14	3

04	6	6	15	3
07	6	6	6	12
06	3	5	17	2
08	5	5	18	1
11	4	5	19	0
17	4	4	20	0
01	5	3	19	2
13	5	2	22	0

Table 7 shows the essay ratings for the G domain. Unlike the RAC and CO domains, articles with higher criterion scores do not tend to receive more accurate ratings from peers. As a result of their inaccurate score distributions, peer raters also overrated rather than underrated for the G domain, indicating that they assigned higher scores than expert raters.

Table 7. Ratings on essays for G domain (ordered from high to low accuracy).

Essay No.	Criterion Scores	Accurate Ratings	Inaccurate Ratings	
			Above	Below
09	5	13	5	6
03	5	12	11	1
14	5	11	7	6
04	6	9	12	3
05	4	9	6	9
11	4	8	15	1
15	6	8	11	5
08	5	7	16	1
07	5	6	16	2
10	6	6	1	17
16	5	6	18	0
12	4	5	17	2
18	5	5	18	1
17	4	4	20	0
06	3	3	12	9
01	5	2	21	1
02	4	1	22	1
13	4	0	24	0

As shown in Table 8, concerning the comments provided, they were classified into four distinct categories: complimentary, neutral, critical, and perplexing. The remarks

from expert raters predominantly focus on delineating performance within the respective mark bands. In contrast, peer raters' comments span a wider range of categories. This variance indicates a discrepancy in perspective between expert raters and peer raters in terms of the specific domain of the article. This incongruity implies that peer raters who diverge from expert raters in their approach might potentially provide inaccurate ratings within the domain. Furthermore, the perplexing comments shed light on the misunderstandings held by peer raters, subsequently leading to the issuance of inaccurate ratings within the particular domain.

Table 8. Sub-themes of peer rating reflections on writing domains.

Writing Domain	Complimentary Comment	Neutral Comment	Critical Comment	Perplexing Comment
RAC	- Highly relevant content - Adequate answer to the task set - Easy to read and understand - A clear and unambiguous viewpoint	- Mediocre performance on content relevance - Arguments could be better developed	- Low relevance of the content - Insufficient argumentation - One-sided argument - Unbalanced argumentation of subpoints - Lack of clarity of viewpoint - The article is hard to understand	-
CO	- Balanced paragraph development - Clearly structured and organised - Paragraphs are properly connected - Concise and clear introduction and conclusion	- Minor problems with paragraph proportions - Minor problems with paragraph organisation - Slightly incoherent paragraphing	- Imbalance in the proportion of paragraphs - Inappropriate organisation of paragraphs - Conclusion does not fulfill the requirements	- Associating it with other domains like RAC, C, V, and G
C	- Satisfactory use of cohesion and cohesive devices - Sequencing ideas rather logically and effectively	- For the most part satisfactory cohesion - Occasional deficiencies cause communication is sometimes ineffective	- Unsatisfactory cohesion - Lack of cohesive devices - Inappropriate use of cohesive devices	- Associating it with other domains like RAC, CO, V, and G

V	- Skillfully uses a wide range of lexical items - Accuracy in word choice and collection	- Uses a rather wide range of vocabulary - Some lexical inappropriacies	- Frequent inadequacies in vocabulary and very limited vocabulary repertoire - Frequent lexical inappropriacies - Inadequate vocabulary for basic communication	- Associating it with other domains like CO, C, and G
G	- Sentences are almost all free of grammatical error - A wide range of grammatical structures	- Grammatical errors are observed occasionally - Complex structures are used with occasional errors - Punctuational errors are observed occasionally - Complex grammatical structures are used occasionally	- Frequent grammatical errors - Limited range of grammatical structures - Complex grammatical structures are often used inaccurately	-Associating it with other domains like CO, C, and V

5 Discussion

For Research Question 1, the detailed results shed light on the accuracy of the 24 peer raters in scoring the various rating domains based on diverse knowledge backgrounds, including sociolinguistics, linguistics, and discourse (Aryadoust, 2012; Weir, 1990). Notably, the RAC domain, which draws upon sociolinguistic knowledge, received the highest level of accurate scoring from the peer raters. This suggests that they were adept at evaluating the appropriateness and adequacy of the content in relation to interactive topics or task settings. Similarly, the V and G domains, which rely on linguistic knowledge, were also scored relatively accurately by the peer raters. However, accurately scoring the CO and C domains, which involve discourse knowledge, presented greater challenges for the peer raters. These domains encompass elements such as the overall structure, organisation, coherence, and cohesion of the writing. The intricacies of assessing these aspects, including the logical flow of ideas and the smooth transition between sentences and paragraphs, contributed to the lower levels of accuracy observed in the peer raters' evaluations. This divergence underscores the necessity for targeted interventions aiming at bolstering students' evaluative judgement, particularly in the domains where deficits were observed.

Regarding Research Question 2, this study investigates the alignment and disparities between peer raters' and expert raters' evaluations of 18 essays across five writing rating domains. The analysis reveals three main findings. Firstly, essays with higher criterion scores are more accurately assessed by peer raters in the RAC and CO dimensions, indicating a positive correlation between domain quality and accurate peer assessment. Second, as in previous peer assessment studies (Esfandiari & Myford, 2013; Farrokhi et al., 2012; Matsuno, 2009; Saito & Fujita, 2004), an interesting pattern emerges where peer raters tended to assign higher scores than criterion scores, suggesting their tendency to overrate rather than underrate essays. Furthermore, expert raters' rating reflections mainly focused on the corresponding performance levels' assessment criteria. In contrast, peer raters' remarks varied between different performance levels and were even confused with assessment criteria from unrelated rating domains. This disparity in justifications for scoring decisions may partly explain inaccurate ratings in specific domains, particularly when perplexing comments reflect misunderstandings among peers.

6 Conclusion

Based on a mixed-methods approach, the study examines the accuracy of peer assessment for ESL argumentative writing as well as the justifications provided by peer raters for their scoring decisions, revealing why some domains are challenging to peer assess accurately, but it is not exhaustive. Its limitations raise the possibility of including process-related data, such as verbal protocols, eye-tracking data, and psychometric data, which could enhance the depth of analysis by revealing the complex cognitive processes underlying peer assessment (Xie et al., 2024). The integration of such data would not only provide a finer-grained understanding but could also potentially offer explanations for the patterns and outcomes observed. This is a promising avenue for future exploration, one that could significantly enhance the robustness of the study's findings.

Acknowledgements. Our thanks go to Kaihua He and Yiming Jing, who provided valuable assistance by crosschecking the thematic codes and comparing the qualitative results that were derived independently. The Ethics Committee for Research Involving Human Subjects (JKEUPM) at Universiti Putra Malaysia granted approval for this study (reference number: JKEUPM-2023-411).

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

References

1. Aryadoust, V.: Evaluating the psychometric quality of an ESL placement test of writing: A many-facets Rasch study. Linguistics Journal **6**(1), 8-33 (2012)
2. Aryadoust, V., Tan, H. A. H., Ng, L. Y.: A Scientometric review of Rasch measurement: The rise and progress of a specialty. Frontiers in Psychology **10**, 2197 (2019)

3. Brunswik, E.: The conceptual framework of psychology. University of Chicago Press, Chicago (1952)
4. Engelhard, G. Jr.: Evaluating rater accuracy in performance assessments. Journal of Educational Measurement **33**, 56–70 (1996)
5. Engelhard, G. Jr.: Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. Routledge, Oxfordshire (2013)
6. Engelhard, G. Jr., Wind, S.: Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge, Oxfordshire (2017)
7. Esfandiari, R., Myford, C. M.: Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. Assessing Writing **18**(2), 111-131 (2013)
8. Falchikov, N., Goldfinch, J.: Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. Review of Educational Research **70**(3), 287-322 (2000)
9. Farrokhi, F., Esfandiari, R., Schaefer, E.: A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. JALT Journal **34**(1), 79-101 (2012)
10. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. International Journal of Qualitative Methods **5**(1), 80-92 (2006)
11. Han, C.: Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. Assessment & Evaluation in Higher Education **43**(6), 979-994 (2018)
12. Han, C., Zhao, X.: Accuracy of peer ratings on the quality of spoken-language interpreting. Assessment & Evaluation in Higher Education **46**(8), 1299-1313 (2021)
13. Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., Suen, H. K.: Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. Assessment & Evaluation in Higher Education **41**(2), 245-264 (2016)
14. Matsuno, S.: Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. Language Testing **26**(1), 75-100 (2009)
15. Saito, H., Fujita, T.: Characteristics and user acceptance of peer rating in EFL writing classrooms. Language Teaching Research **8**(1), 31-54 (2004)
16. Topping, K.: Peer assessment between students in colleges and universities. Review of Educational Research **68**(3), 249-276 (1998)
17. Wang, J., Engelhard, G. Jr.: Conceptualizing rater judgments and rating processes for rater-mediated assessments. Journal of Educational Measurement **56**(3), 582-609 (2019)
18. Wang, J., Engelhard, G. Jr., Raczynski, K., Song, T., Wolfe, E. W.: Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. Assessing Writing **33**, 36-47 (2017)
19. Weir, C.: Communicative language testing. Prentice-Hall, London (1990)
20. Wind, S. A., Engelhard, G. Jr.: How invariant and accurate are domain ratings in writing assessment?. Assessing Writing **18**(4), 278-299 (2013)
21. Xie, X., Nimehchisalem, V., Yong, M. F., Yap, N. T.: Malaysian students' perceptions towards using peer feedback to cultivate evaluative judgement of argumentative writing. Arab World English Journal **15**(1), 298-313 (2024)