

Determining English Language Lecturers' Quality of Marking in Continuous Assessment through Rasch Analysis

Mardiana Idris

Education Department, Institut Pendidikan Guru Kampus Temenggong Ibrahim, Jalan Datin Halimah, 80350, Johor Bahru, Johor, Malaysia

ABSTRACT

English language lecturers at matriculation colleges are generally equipped with assessment criteria for marking students' written assessment. However, these criteria are normally susceptible to lecturers' interpretation and understanding, which threatens quality marking. Therefore, this study aimed to determine the severity and consistency of English language lecturers' marking of English academic writing (EAW) in continuous assessment. The participants were five English language lecturers and 50 matriculation students. Each lecturer selected ten EAWs randomly from 318 matriculation students. The five-part EAW was marked first by the class's lecturer and later, it was marked by pre-assigned peer moderators who did not teach the students. The total data set collected was 250 (5 lecturers x 10 EAWs x 5 parts of EAW). The data were analyzed with Many-Facets Rasch Measurement (MFRM) application. Semi-structured interviews were conducted with both lecturers and students for triangulation purposes. Findings revealed that four out of five lecturers were lenient in marking but the marking was found to be internally consistent with infit and outfit mean squares for each lecturer ranged between 0.5 and 1.5. From interview responses analyzed, students perceived their lecturers as fair but strict in awarding marks. These responses were consistent with most lecturers' responses on their strict adherence to assessment criteria. Discussion of findings is centered on the issue of severity and consistency of the assessors. This study could offer a practical solution in

providing evidence for quality marking of written assessment and, consequently, aid in developing remedial measures for misfit assessors in educational institutions.

ARTICLE INFO

Article history:

Received: 16 July 2021

Accepted: 04 October 2021

Published: 30 November 2021

DOI: <https://doi.org/10.47836/pjssh.29.S3.19>

E-mail address:

mardiana.idris@ipgm.edu.my

Keywords: Consistency, continuous assessment, Rasch analysis, severity, written assessment

INTRODUCTION

Generally, continuous assessment is perceived as a measurement mechanism to gauge the learners' learning progress and gain based on specified and fixed criteria, which normally translate learners' achievement into numerical digits (Carrillo-de-la-Pena & Perez, 2012; Mikre, 2010; Walde, 2016). These digits are then converted into grades, bands, categories, or levels that portray learners' ability to master skills, topics, or subjects. However, how accurate is this portrayal, particularly when it involves subjective marking whereby the assessors solely awarded marks? Despite each assessor's every intention to remain objective, to compound the conundrum further, their marking may be 'affected by classroom relationships and interactions' (Tierney, 2016) in the teaching and learning environment. It leads to the issue of ensuring quality in marking. Quality marking is essential, particularly in continuous assessment, because it affects students' learning. Tierney (2016) and Jiminez (2015) reported that learners generally exhibited their actual performance in learning if they perceived the teachers or lecturers as being fair in assessing their assessments. Therefore, this study attempted to determine lecturers' severity and consistency of marking matriculation English academic writing (EAW) in a continuous assessment.

In this paper, the objective and research questions are first outlined. Then, theoretical underpinnings of assessment and studies related to severity and consistency in marking are discussed in the literature

review. Subsequently, the methodology used is elaborated, and this is followed by describing the findings based on the research questions. Finally, discussion, implications, and conclusions are presented.

OBJECTIVE OF THE STUDY

The study's primary objective was to determine English language lecturers' severity and consistency in marking matriculation students' five-part English academic writing (EAW) paper.

RESEARCH QUESTIONS

Three research questions were formulated to guide the study to achieve the primary objective

1. to what extent were the assessors severe in marking matriculation students' EAW in continuous assessment?
2. to what extent were the assessors consistent in marking matriculation students' EAW in continuous assessment?
3. how did lecturers and students perceive the severity and consistency of EAW marking in continuous assessment?

LITERATURE REVIEW

Severity and leniency in marking written assessments have always been dilemmas faced by many lecturers or assessors. Questions that linger include "Did I mark according to the rubric provided?", "Did I award an 'accurate' score that reflects the student's performance?" and "Did my

assumptions of the students' knowledge or behavior cloud my fair judgement?" These lingering quality control indicators may have resulted in learners questioning the scores or marks they have received, particularly if they perceived that they had been assessed severely or unfairly by their assessors. Assessor or rater severity consistently provides scores or 'ratings that are lower or higher than is warranted' (Engelhard, 1994) by learners' performances. In fact, there are many studies on severity of assessors (Han & Huang, 2017; He, 2019; McNamara et al., 2019; Park, 2011) in assessing written task and its impact on quality assessment. Levey (2020) observed that any performance assessment typically judged by human raters will introduce subjectivity. Consequently, this could lead to unreliable scoring. Studies by Fahim and Bijani (2011) as well as Erguvan and Dunyait (2020) reported that assessors' severity and leniency in marking could cause dissatisfaction among test takers, and both studies recommended for rater training to be given to assessors in order to reduce rater variability. Most studies reported that rater training did reduce rater variability but did not eliminate it.

Another imperative criterion for quality marking is consistency, which is often linked to reliability. This study obtained assessors' consistency by providing training for assessors and using multiple assessors (Lang & Wilkerson, 2008; Willey & Gardner, 2010). Many studies have reported the importance of training the assessors before marking to achieve a higher inter-rater or consistency value (Erguvan & Dunyait, 2020; Kayapinar, 2014; Park,

2011; Sundqvist et al., 2020). For example, Hack (2019), in her doctoral thesis on marking processes used in the assessment of extended written responses, quoted a study by Morin et al. (2018) which reported that 'the probability that candidates receive the correct grade (the 'definitive' grade awarded by the team of senior examiners) on a combined English literature and language qualification was only 52%.' (p. 10). Thus, this indicates that the reliability of marking written assessment invites contention if not conducted properly.

Emphasis on the severity and consistency of marking is due to its feedback role in the formative assessment framework. Black and William (2009) conceptualized five key strategies in the assessment process. The first strategy was to clarify and share learning intentions and criteria. The second strategy involved engineering learning tasks that elicit evidence of student learning. Finally, the third strategy pertained to providing feedback that moves learners forward. It was achieved through written feedback given by fair and consistent markers, which guided their subsequent performance. The fourth strategy concerned activating learners as instructional resources, while the fifth focused on activating learner, as the owners of their learning. The framework for assessment strategies is illustrated in Table 1.

The conceptualized framework by Black and William (2009) in Table 1 shows that assessment contributes to quality learning. A direct consequence for learners' improvement in writing skills is through column 3, "Providing feedback that moves

Table 1
Assessment strategies framework suggested by Black and William (2009)

	Where the learner is going	Where the learner is right now	How to get there
Teacher	1. Clarifying learning intentions and criteria for success	2. Engineering effective classroom discussions and other learning tasks that elicit evidence of student learning	3. Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	4. Activating learners as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	5. Activating learners as the owners of their learning	

learner forward.” Hypothetically, suppose students received unfair and inconsistent marks or scores as feedback for their written assessment. In that case, it could indirectly affect their learning because feedback or scores given does not truly reflect their ability. As such, learners ‘may be moved’ in the wrong direction in improving their writing skills.

The severity and consistency of assessors could always be gauged through classical test theory, whereby average scores and reliability of assessors are analyzed. However, this theory alone is not enough to describe the linear relationship between students, items, and subjective marking of assessors. Hence, Many-Facets Rasch Measurement (MFRM) was used in this study. MFRM is a psychometric analysis that can identify assessors’ severity and consistency in marking subjective assessment (Prieto & Nieto, 2014; Eckes, 2005). Meadows and Billington (2005) outlined the advantages of MFRM, which include:

“Using a many-facets analysis, each question paper item or behavior that

was rated can be directly compared.

In addition, the difficulty of each item, as well as the severity of all judges who rated the items, can also be directly compared. Person abilities can be evaluated whilst controlling for differences in item difficulty and judge severity.” (Meadows & Billington, 2005; p. 6)

Based on these advantages, the MFRM has been used in many large-scale assessments and certifications, including developing the Common European Framework of Reference (CEFR) (Council of Europe, 2009).

METHODOLOGY

This methodology section describes the participants involved in the study and the instruments used to collect the data. The nine phases of the procedures are also described.

Participants

The lecturers (labeled as assessors henceforth) were five English language

lecturers who taught matriculation English 1, English 2 and Malaysian University English Test (MUET) to matriculation students. The assessors had ten to fourteen years of teaching experience. Four out of five assessors had experience in marking the MUET Writing paper. In addition, all assessors were well versed with the rubrics and scoring guide provided by the Matriculation Division as they had been given training prior to marking the assessment. Based on the appointment letters by the college, each lecturer was appointed as an assessor for their own students' assessment and a moderator for their peers. One of the lecturers was appointed as a chief moderator.

As for students, they were 50 engineering matriculation students. On average, they were 18 years old. Most students were categorized as having intermediate to advanced levels of English language proficiency based on their *Sijil Pelajaran Malaysia* (SPM) English results.

Instruments

Two types of instruments were used in this study—students' EAW and a semi-structured interview. Fifty EAWs were randomly selected from 318 matriculation students. The 50 scripts were selected due to the procedures outlined by the Matriculation Division, whereby English language lecturers must moderate ten EAW scripts from their classes. For the EAW, the students were required to write a personal statement to a university for placement purposes. Students had to write their statements in

five parts. Part 1 was an introduction to the personal statement. Part 2 was a content paragraph in which students were required to describe their past experiences using the past tense. Part 3 was another content paragraph that required students to describe their current undertakings, while Part 4 was the last content paragraph which required students to write in the future tense. Finally, Part 5 was the conclusion to the personal statement. For a complete sample of the paper, please refer to Appendix A.

In terms of scoring criteria, Part 1 and 5 used five scoring levels, with Level 1 (Limited user) as the lowest and Level 5 (Excellent user) as the highest. Generally, Parts 1 and 5 employed holistic assessment criteria (Appendix B). As for the content paragraph, it also used five scoring levels. The levels were: Level 1 (very weak), Level 2 (weak), Level 3 (Fair), Level 4 (Good), and Level 5 (Very Good). However, Parts 2, 3, and 4 used an analytic assessment criterion that focused on three components: focus, organization, and language (Appendix C)

Semi-structured interviews with lecturers and students were also conducted to corroborate the quantitative findings.

Procedures

The study was conducted in nine phases. Phase 1 focused on training the assessors and the moderators to mark the EAWs. Chief Moderator gave the training, and during training, assessors were encouraged to ask questions to have the same understanding of the criteria. After all, assessors were clear with the rubrics and scoring guide,

and they conducted the same briefing to their students prior to assessment. Next, students attempted EAW in Phase 2. Every assessor marked their scripts for two weeks in Phase 3. Then, in Phase 4, scripts were moderated by peer moderators. For Phase 5, none of the scripts had to be moderated by the Chief Moderator since the difference in raw scores was not more than five marks. Phases 6, 7, and 8 involved MFRM analysis, interviews, and transcription. Finally,

Phase 9 concentrated on the findings. The summary of all nine phases involved is presented in Table 2.

FINDINGS

Descriptive statistics, Rasch variable map (Wright map), assessor measurement report, and interview responses are used to report the findings based on the research questions initially presented.

Table 2
Summary of nine phases of the study

Phase	Description	Analyses involved
Phase 1	<ul style="list-style-type: none"> Lecturers were appointed as assessors for the continuous assessment. Assessors received training on scoring guides and criteria from the Chief Moderator. Assessors asked questions to the Chief Moderator when doubts arose. All matriculation students were given the scoring guide and criteria. Lecturers explained the scoring guide and criteria to the students. 	Not applicable
Phase 2	<ul style="list-style-type: none"> 318 students attempted all five parts of the EAW. 	Not applicable
Phase 3	<ul style="list-style-type: none"> Each assessor randomly selected 10 EAW to be marked using the scoring guide and criteria. Assessors were given two weeks for marking. 	Raw scores
Phase 4	<ul style="list-style-type: none"> Each assessor submitted their ten (10) marked EAW scripts to their peer moderator. Moderators were given a week to mark. Rating/judging designs for both assessors and moderators were preplanned to ensure a smooth analysis in the MFRM software (Facets) 	Raw scores
Phase 5	<ul style="list-style-type: none"> Moderators returned the marked scripts to the first assessors. Since the difference of marks was not more than five in each EAW, the scripts were not submitted to the Chief Moderator. 	Raw scores
Phase 6	<ul style="list-style-type: none"> The researcher analyzed the data in Facets software: 3 facets rating scale—assessors, students' EAW, and items with rating 1 to 5. 	Facets analysis
Phase 7	<ul style="list-style-type: none"> A semi-structured interview was conducted with lecturers. A semi-structured interview was conducted with students. 	Not applicable
Phase 8	<ul style="list-style-type: none"> Transcription of interview 	Thematic analysis
Phase 9	<ul style="list-style-type: none"> Analysis of findings 	<ul style="list-style-type: none"> Descriptive statistics Rasch variable map (Wright Map) Assessor measurement report

Descriptive Statistics

Table 3 shows mean ratings by lecturers for parts 1, 2, 3, 4, and 5 of EAW. Based on Table 3, it shows that Lecturer 5 seemed to be severe with the rating awarded as the mean for each part was categorized as a competent user (3) and fair (3) while the rest of the lecturers were awarded good standing (4) for most parts of the EAW. At first glance, it could indicate that Lecturer 5 was severe in marking, but this did not entirely explain the severity of the assessor since it was based on means. Therefore, MFRM analysis was used.

Severity of Assessors in Marking EAW

Figure 1 illustrates a graphical description of three facets analyzed in the MFRM – student ability, part (or item) difficulty, and assessor severity- along a logit scale of a Rasch ruler. Logit is the unit used in reporting the MFRM analyses. The first column is a measure column (Measr) which ranges between -2 logits and +8 logits, with 0 as the mean. The second column (Students) displays students’ ability based on the ratings awarded. Higher ability students are closer to the top, while less able students are closer to

the bottom. The third column displays the five parts of the EAW. The parts are ordered according to the level of severity imposed by assessors. The harsher a part is assessed,

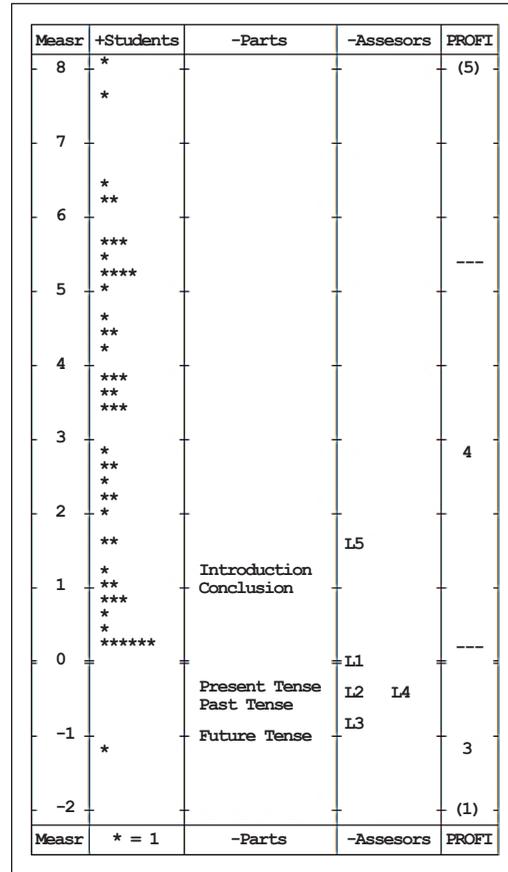


Figure 1. The Wright map for students’ ability, level of EAW difficulty, and assessors severity

Table 3
Mean rating by lecturers for parts 1, 2, 3, 4, and 5 of EAW

Lecturer/ Part	Lecturer 1	Lecturer 2	Lecturer 3	Lecturer 4	Lecturer 5	Mean
1 (Introduction)	3.5	3.9	4.2	4.3	2.9	3.7
2 (Past Tense)	4.1	4.4	4.6	4.3	3.6	4.2
3 (Present Tense)	4.1	4.3	4.5	4.4	3.6	4.2
4 (Future Tense)	4.0	4.6	4.8	4.4	3.6	4.3
5 (Conclusion)	3.6	3.7	4.0	4.3	3.5	3.8
Mean	3.9	4.2	4.4	4.3	3.4	4.04

the higher is the position of the part on the map. Conversely, the lower the position of a part on the map, the less harsh the part is assessed. The fourth column displays five assessors coded as L1 to L5. Severe assessors are located closer to the top, while lenient assessors are located closer to the bottom. The fifth column displays the rating scale used (1– 5).

Based on Figure 1, the student ability ruler indicates that the students scored highly on the EAW as 49 out of 50 students were above mean 0 while only one student was rated below mean 0. In addition, student ability was clustered within scale 4 (good) as indicated from 0 logits to +8 logits. This distribution pattern implied that students could be highly proficient despite being randomly selected by the lecturers.

Next to the student ability column is the part ruler. The parts are ordered with an introduction as the harshest part rated by assessors while future tense as least harshly rated. There seems to be a clear pattern distinction as the introduction and conclusion (holistic criteria) are closer together. In contrast, present, past, and future tense (analytic criteria) are clustered together. Despite this distinction, the parts do not differ much within -1 logit and 1 logit. It suggests that both analytic and holistic criteria received approximately similar attention from the assessors since they are clustered together. However, holistic criteria (Introduction and Conclusion) seem to receive more attention than analytic criteria since they significantly differ from the rest.

Besides parts, assessors are also modeled with the most severe ones at the

top and the most lenient ones at the bottom of the Rasch ruler. The ruler shows that L5 is the most severe assessor while L3 is the most lenient. The map also indicates more lenient raters than severe ones as four assessors are positioned below mean 0.

The final ruler displays the five rating scales. The range of the rating ruler for all five categories starts from 1 until 5. Although the rating scale has five levels, levels 1 and 2 are absent from the ruler. It implies that these levels were not awarded to students.

Consistency of Assessors in Marking EAW

The Wright map described earlier was only a brief representation of all the facets investigated for quality control. Therefore, to address the second research question, an assessor measurement report is needed.

Table 4 shows the assessor measurement report, ordered from the most severe to the most lenient raters. Infit and outfit mean-squares for four raters were between 0.5 and 1.5 logits, and these values were the recommended range for productive measurement. Although the infit and outfit mean-squares of L2 (infit: 1.58 and outfit 1.55) slightly exceeded the recommended range, these values, however, did not distort the measurement as they did not exceed 2.0. According to Linacre (2014), separation of more than two and reliability of more than 0.8 were indications of data that fit the measurement model. The values of separation and reliability statistics provided at the bottom of Table 4 indicated that the data fitted the model since the

Table 4
Measurement report on lecturers' severity in marking

Lecturers	Severity Measure	Model S.E	Infit MnSq	Outfit MnSq
L5	1.66	0.18	0.41	0.39
L1	-0.01	0.20	1.14	1.24
L4	-0.37	0.23	1.13	1.18
L2	-0.44	0.22	1.58	1.55
L3	-0.84	0.24	0.88	0.85

Note. S.D: 0.95; Separation: 4.44; Reliability (not inter-rater): 0.95

separation was 4.44, and thus, reliability was high with 0.95. In addition, the standard deviation (S.D) given at the bottom of the Table 4 indicated that the data were clustered towards the mean with less than one standard deviation. It suggests that assessors had a similar rating tendency. As for assessor severity, this was gauged from logit measures reported in the second column of the table. The range of severity measure from the most severe assessor (L5: 1.66 logits) to the most lenient (L3: -0.84 logits) was about 2.5 logits. Table 4 shows that four out of five assessors were lenient in awarding their ratings for the written assessment.

From the severity measures provided, it was found that most assessors tended to rate the essays leniently. However, the severity measures of L1, L2, L3, and L4 did not differ much, and most importantly, they did not exceed -1 logits. In fact, since the severity measures clustered between -0.01 logits and -0.84 logits, it might indicate that they had a similar understanding of the assessment criteria. However, the L5 severity measure exceeded 1 logit (1.66). Therefore, it may indicate a departure from applying the assessment criteria objectively.

Internal consistency was measured through assessors' infit mean-squares. Infit mean-square is less sensitive to outliers, but they are more sensitive towards unexpected ratings (Yan, 2014). Hence, infit mean-square is the benchmark for assessors' internal consistency in awarding scores. Based on Table 4, L5 displays infit mean-squares lower than 0.5 (0.41 logits), which indicated that the value was influenced by rating patterns and thus, posed a greater threat to measurement (Linacre, 2014). Although the L2 infit mean-square was 1.58 logits, this value did not distort the measurement as it did not exceed 2.0 logits. The infit mean-squares of three assessors were between 0.88 logits and 1.14 logits. These values indicated that most assessors were largely internally consistent in marking the EAW.

Perception on Severity and Consistency of Marking EAW by Assessors and Students

Analyses from the semi-structured interviews revealed a stark contrast between what was perceived by the students and the lecturers with the MFRM analysis obtained. Two questions were posed to students:

1. Do you think your lecturer was fair in marking your essays? Please provide your reasons.
2. Do you think your lecturer was strict in awarding you the marks? Please state your reasons.

For the first question, all the students believed their lecturers were fair in awarding the EAW marks. Two themes emerged from their reasoning: 1) marks awarded reflected students' performance or ability, and 2) marks awarded the assessment criteria. More than half of the students mentioned that the marks awarded were based on their performance in writing, and therefore, they perceived it as fair. For example, S2 remarked that "because it depends on my writing task. She knows how to evaluate it," while S23 justified the marks given by stating (verbatim), "I can see which task my weakness and the marks are given is what I deserve." Nearly half of the students also opined that their lecturers assessed their EAW based on the assessment criteria. For example, S1 justified the marks received by stating, "I know my lecturer gave it by following the guidelines." At the same time, S20 observed that "I think everyone is treated fairly according to the rubric."

As for the second question, most students believed their lecturers were strict in awarding them the marks. Only two students (S11 and S21) were not sure whether their lecturer (L3) was strict in awarding them marks, while five students (S1, S4, S18, S20, and S22) thought that their lecturers (L2, L3, and L4) were not

strict in awarding marks. Most students, justified their reasoning positively despite stating that their lecturers were strict in awarding marks. For example, S6 remarked that "I did not get a very high mark but get the marks that equivalent to what I do," and S12 concurred by claiming that "because she gives the marks follow by student's talented (skills)." S9 believed that his lecturer had to be strict because "she needs to do so to make sure all her students were excellent."

When questions on severity and consistency of marking were directed towards the lecturers, most lecturers maintained that they would not be strict unnecessarily as they followed the assessment criteria closely while marking the written assessment. It is evident from their responses:

L1: *"I'm not strict in awarding the marks but at the same time I would follow the assessment criteria closely. I will not penalize the marks unnecessary."*

L5: *"Scripts were assessed on fluency, organization and language. Therefore, being strict is a subjective connotation."*

As for consistency, most lecturers believed that they were consistent in their marking as illustrated by the reasoning given by L1 ("I will compare the marks with other scripts if I have any doubt with the marks that I have awarded") and L3 ("I follow the criteria while marking and it is always in front of me").

DISCUSSIONS AND IMPLICATIONS

Findings from the MFRM analysis indicated that only one lecturer was more severe than others (L5: 1.66 logits on severity measure). In contrast, most students perceived that their lecturers were severe or strict in awarding marks, albeit accompanied by positive reasons for why they deserved the marks. This finding is consistent with studies by Fahim and Bijani (2011), and Erguvan and Duniyait (2020), which found that despite training provided, assessors' severity and harshness could not be eliminated. In addition, both students and lecturers were generally unanimous in their perceptions of assessment fairness. This could be attributed to the fact that both parties were exposed to the scoring guide and criteria at the onset of the study (Phase 1). Their responses mirrored Nisbet and Shaw's (2020) 'felt fairness.' In their book 'Is Assessment Fair?' they argued that a sense of fairness carries 'emotive force' and thus, any perception towards fairness in assessment deserves attention. In fact, they highlighted the challenges in 'harmonizing' other assessment concepts, such as validity and reliability with assessment fairness. Since fairness is subjective, students and lecturers' responses in this study were valuable. They provided a glimpse of how quantitative and qualitative findings could offer an inclusive view of assessment concepts.

Consistency or reliability of marking is important in ensuring quality marking. This study indicated that most lecturers were reliable markers based on their infit

mean-squares—ranged between 0.5 and 1.5. In addition to the training provided, it could be hypothesized that their experience in marking standardized examination papers like the MUET might have helped them internalize the assessment criteria. In this study, only L5 (infit mean square: 0.41 logits) did not have extensive experience in marking compared to the rest of the lecturers. However, L5's lack of internal reliability should not be construed as the failure of training given. Other factors could affect the reliability of markers, such as rater fatigue (Mahshanian & Shahnazari, 2020).

Based on the discussion of findings, this study offers a two-pronged solution to two assessment concerns. The first concern pertains to producing evidence of quality marking of written assessment, and the second is to diagnose misfit assessors for remedial measures. Providing a quality rubric does not necessarily translate to quality marking as its application or interpretation may get lost in translation. Therefore, using statistical analyses such as MFRM may provide evidence of quality marking. Educational institutions could download the free version of MFRM (Minifac), which enables its user to analyze up to 2000 data (Linacre, 2014).

From the MFRM measurement reports, misfit assessors could be identified, and remedial measures could be taken. For example, more training and moderation exercises could be prepared for assessors who exhibit variability in marking. Assessor variability could not be eliminated in any performance assessment. However, by

devising appropriate measures to control the marking quality, students will receive fair and just marks or scores that correspond with their ability.

CONCLUSIONS

Many studies on severity and consistency of raters in marking written assessment reported that rater training was crucial in maintaining quality marking. (Park, 2011; Han & Huang, 2017; He, 2019, McNamara et al., 2019). The findings of this study seemed to corroborate this stance as most lecturers were able to mark after training was provided reliably. In addition, the utilization of the MFRM in gauging severity and consistency measures of assessors' tendency in marking contributed to the burgeoning literature of performance assessment. The availability of psychometric testing software such as MFRM enables educational institutions to portray quality marking accurately. Triangulation between Rasch analyses and students' and lecturers' interview responses produced interesting insight into assessment fairness. Fairness has always been a persistent contention in any performance assessment, and hopefully, this finding could add value to its literature.

There were some limitations identified in this study. Firstly, it was found that despite the EAW being randomly selected, the students' scores revealed that most of them were categorized as proficient. This could affect their perception of fairness since the marks were in their favor. It would be ideal to employ students with varying proficiency levels (beginner, intermediate

and advanced) in future studies and then interview them on their perception of fairness. Secondly, there were only five lecturers involved in this study. Despite obtaining sufficient data points for MFRM analysis, using a bigger number of lecturers might yield different results in terms of severity and leniency measures. Thirdly, the training provided in this study was short due to lecturers' work commitment. Thus, future studies may want to include longer training hours in their procedures, particularly for novice assessors.

ACKNOWLEDGEMENT

The author wants to extend utmost gratitude to former matriculation college Director Azman bin Mokhtar and the Matriculation Division for their support and permission to publish this paper. The author is also grateful for the constructive comments received from the anonymous reviewers and the editors from *Pertanika Journal of Social Sciences and Humanities* (Special Issue).

REFERENCES

- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31. <https://doi.org/10.1007/s11092-008-9068-5>
- Carrillo-de-la-Pena, M. T., & Perez, J. (2012). Continuous assessment improved academic achievement and satisfaction of psychology students in Spain. *Teaching of Psychology*, 39(1), 45-47. <https://doi.org/10.1177/0098628311430312>
- Council of Europe. (2009). *Common European Framework of Reference for Languages:*

- Learning, teaching, assessment*. Cambridge University Press.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2), 93-112. <https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Erguvan, I. D., & Dunya, B. A. (2020). Analyzing rater severity in a freshman composition course using Many-Facet Rasch measurement. *Language Testing in Asia*, 10(1), 1-20. <https://doi.org/10.1186/s40468-020-0098-3>
- Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, 1(1), 1-16.
- Hack, S. (2019). *How do examiners mark? An investigation of marking processes used in the assessment of extended written responses* [Unpublished Doctoral dissertation]. University of Surrey.
- Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, 53, 112-147.
- He, T. (2019). The impact of computers on marking behaviors and assessment: A many-facet Rasch measurement analysis of essays by EFL college students. *SAGE Open*, 9(2), 1-17. <https://doi.org/10.1177/2158244019846692>
- Jimenez, C. E. (2015). *Middle school students' perceptions of fairness and trust in assessment scenarios* (Doctoral dissertation). University of South Carolina, US.
- Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136. <https://doi.org/10.14689/ejer.2014.57.2>
- Lang, W. S., & Wilkerson, J. R. (2008, February 7-10). *Accuracy vs. validity, consistency vs. reliability, and fairness vs. absence of bias: A call for quality*. Paper presented at the Annual Meeting of the American Association of Colleges of Teacher Education (AACTE). New Orleans, LA.
- Levey, D. (2020). *Strategies and analyses of language and communication in multilingual and international context*. Cambridge Scholars Publishing.
- Linacre, J. M. (2014). *A user guide to Facets, Rasch-model computer programs*. Winsteps.com
- Mahshanian, A., & Shahnazari, M. (2020). The effect of raters fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics*, 10(1), 1-13. <https://doi.org/10.17509/ijal.v10i1.24956>
- McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment - Oxford applied linguistics*. Oxford University Press.
- Meadows, M., & Billington, L. (2005). *A review of the literature in marking reliability*. National Assessment Agency.
- Mikre, F. (2010). The roles of assessment in curriculum practice and enhancement of learning. *Ethiopian Journal of Education and Sciences*, 5(2), 101-114. <https://doi.org/10.4314/ejesc.v5i2.65376>
- Morin, C., Black, B., Howard, E., & Holmes, S. D. (2018) *A study of hard-to-mark responses: Why is there low mark agreement on some responses?* Ofqual Publishing.
- Nisbet, I., & Shaw, S. (2020). *Is assessment fair?* SAGE Publications Ltd.
- Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory*

- and item response theory* [Doctoral dissertation]. Columbia University.
- Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many-Faceted Rasch Measurement. *Psicológica*, 35, 385-397.
- Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology*, 8(10), 3-29. <https://doi.org/10.46364/njmlm.v8i1.605>
- Tierney, R. D. (2016). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of Educational Philosophy and Theory* (pp. 1-6). Springer Science+Business Media. https://doi.org/10.1007/978-981-287-532-7_400-1
- Walde, G. S. (2016). Assessment of the implementation of continuous assessment: The case of METTU University. *European Journal of Science and Mathematics Education*, 4(4), 534-544. <https://doi.org/10.30935/scimath/9492>
- Willey, K., & Gardner, A. (2010, November 18-19). *Improving the standard and consistency of multi-tutor grading in large classes* [Paper presented]. ATN Assessment Conference 2010. University of Technology Sydney, Australia.
- Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, 31(4), 501-527. <https://doi.org/10.1177/0265532214536171CES>

APPENDICES

Appendix A

A sample essay question on personal statement



PROGRAMME OUTLINE

The programme will provide student with a broad knowledge and in-depth understanding of the following:

- Fundamental of engineering technology in rail transportation
- Rail transportation technology
- Rail transportation safety
- Rail transportation design
- Field Instrument and control in rail transportation

PROGRAMME EDUCATIONAL OBJECTIVES

Knowledgeable in engineering technology in rail transportation discipline in-line with the industry

Technically competent in activities related to engineering technology in rail transportation

Effective in communication with related professionals and stakeholders

Able to adapt in new development related to engineering technology in rail transportation environment

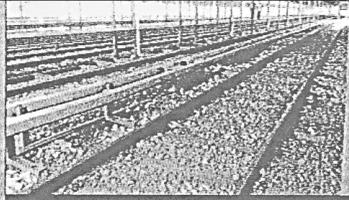
PROGRAMME STRENGTH

Hands-on Program

- 54% Practical, 46% Theory

High industry exposure

- Field Trip
- Work-Based Learning Programme
- Industrial Attachment
- Industrial Lecture Series/Talk



CAREER DESTINATION

- Engineering (CRSI, Infra, Signalling, Electrification, Purchasing, Project)
- Academician
- Railway Operation Management
- Technical and Production Engineer
- Fleet Maintenance
- Project Management
- Executives (Rolling stock, Signalling, Operations, Technical)
- Corporate Strategy
- Permanent Way
- Design Engineer
- Maintenance Engineer
- Technical Support
- Research Officer (R&D)

ENTRY REQUIREMENTS

Sijil Tinggi Persekolahan Malaysia (STPM)

Pass STPM with at least:

- Grade C (NGMP 2.00) in Mathematics T/ Further Mathematics and Physics OR Mathematics T/ Further Mathematics and Chemistry/ Biology (with at least grade C in SPM Physics)

Matriculation (Science/Engineering/Technical)

Pass matriculation with at least:

- Grade C (2.00) in Mathematics/Engineering Mathematics and Physics/Engineering Physics OR Mathematics/ Engineering Mathematics and Chemistry/Biology/Engineering Chemistry/ Engineering Biology (with at least grade C in SPM Physics)

Diploma (Engineering/Technology)

Pass diploma with at least:

- CPA/PNGK 2.50 OR at least CPA/PNGK 2.30 with at least 2 years working experience OR pass Diploma Kemahiran Malaysia (DKM) in Civil Engineering/Mechanical Engineering/ Electrical and Electronics Engineering/ Technology with at least CPA/PNGK 3.00 / Grade B / >80% marks

Other Requirements

- At least Band 2 in Malaysian University English Test (MUET)
- Not blind nor disable person (blind / deaf / lame / mute) which makes it difficult to do practical work

PROGRAMME STRUCTURE

Bachelor of Engineering Technology in Rail Transportation with Honours.
Duration: 4 years

YEAR/SEM	YEAR 1	YEAR 2	YEAR 3	YEAR 4
SEM 1	Academic	Academic	Academic	Academic
SEM 2	Academic	Academic	Academic	Industrial Training
SEM 3		WBL	WBL	

* WBL = Work-Based Learning

UNIVERSITY - INDUSTRY - INSTITUTIONS NETWORKING



PART	ELEMENT	QUESTION
1	Introductory paragraph	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write an introductory paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
2	Body paragraph 1 (past tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
3	Body paragraph 1 (present tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your own personal experience.
4	Body paragraph 1 (future tense)	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a body paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.
5	Conclusion	You are applying for admission to the Bachelor of Engineering Technology in Rail Transportation course at UTHM. Write a conclusion paragraph based on the entry requirements. You may use the vocabulary provided in the visual. You may add your personal experience.

Appendix B

Holistic criteria for Parts 1 (Introduction) and 5 (Conclusion)

Component	5 (Excellent user)	4 (Good user)	3 (Competent user)	2 (Modest user)	1 (Limited user)
PART 1 Introductory paragraph	Excellent developed introductory paragraph. Ideas are excellently linked.	Well-developed introductory paragraph. Ideas are well linked.	Satisfactorily well-developed introductory paragraph. Ideas are satisfactorily linked.	Some development for the introductory paragraph. Ideas are poorly linked.	Limited development for the introductory paragraph. Ideas are not linked.
PART 5 Concluding paragraph	Conclusion summarizes the main topics without repeating previous sentences.	Conclusion summarizes the main topics with minimal repeated sentences.	Conclusion summarizes the main topics with some repeated sentences.	Conclusion summarizes the main topics poorly, repeating previous sentences.	Ends abruptly or no conclusion given.

Appendix C

Analytic criteria for Parts 2, 3 and 4 (Content Paragraphs)

Component (PART 1,2,3)	5 (Very good)	4 (Good)	3 (Fair)	2 (Weak)	1 (Very weak)
FOCUS	Takes a clear position and supports it consistently with well-chosen reasons and/or examples; may use strategies to promote oneself.	Takes a clear position and supports it with relevant reasons and/or examples through much of the paragraphs.	Takes a clear position and supports it with some relevant reasons and/or examples; there are some developments in paragraphs.	Takes a position and provides uneven support; may lack development in parts or repetitive OR paragraphs are no more than a well-written beginning.	Attempts to take a position but the position is very unclear OR takes a position but provides minimal or no support, may only paraphrase the prompt.
ORGANISATION	It is focused and well organized with effective use of transitions.	It is well organized but may lack some transitions.	It is generally organized but has few or no transitions in parts of the paragraphs.	It is disorganized in parts of the paragraph; other parts are disjointed and/ lack transitions.	It is disorganized or unfocused in much of the paragraphs OR is clear but too brief.
LANGUAGE	Consistently exhibits variety in sentence structure and word choice. Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.	Exhibits some variety in sentence structure and uses good word choice; occasionally, words may be used inaccurately. Errors in grammar, spelling, and punctuation do not interfere with understanding.	Most sentences are well constructed but have a similar structure; word choice lacks variety or flair. Errors in grammar, spelling, and punctuation but do not interfere with understanding.	Sentence structure may be simple and unvaried; word choice is mostly accurate. Errors in grammar and spelling and punctuation sometimes interfere with understanding.	Sentences lack formal structure, and the word is often inaccurate. Errors in grammar and spelling and punctuation interfere with understanding in much of the paragraphs.