

RATER CONSISTENCY IN SCORING THE 2024 WASSCE MATHEMATICS ESSAY ITEMS: A MANY-FACET RASCH MEASUREMENT ANALYSIS

Michael Osimhen, PhD

*Dept of Educational Evaluation and Counselling Psychology, Faculty of Education
University of Benin, Edo State, Nigeria.
osimhenmichael@gmail.com
+2348064549081*

&

Iberedem Theresa Ikhiakonotse, PhD

*Dept of Guidance & Counselling, Faculty of Education
Ambrose Alli University, Ekpoma, Edo State.
tessyonotse@aauekpoma.edu.ng
+2348066453562*

Abstract

Rater-mediated scoring of essay items in large-scale examinations is susceptible to measurement error arising from rater severity, leniency, and inconsistency. This study applied the Many-Facet Rasch Measurement (MFRM) model to examine rater consistency in the scoring of the 2024 West African Senior School Certificate Examination (WASSCE) Mathematics essay items. A quantitative ex post facto research design was adopted. The study utilized a Mathematics essay scoring rubric developed in line with WASSCE marking guidelines. The instrument was validated through expert review by three specialists in Mathematics Education and Measurement and Evaluation, and item-level and scale-level Content Validity Indices (I-CVI and S-CVI) were computed, yielding an average S-CVI of 0.87, indicating adequate content validity. Responses from 75 Senior Secondary School III students were selected using purposive sampling, while 15 Mathematics teachers were selected through stratified random sampling from public secondary schools in Alimosho Local Government Area of Lagos State, Nigeria, to serve as raters. Data were analysed using the MFRM model to simultaneously estimate examinee ability, item difficulty, and rater severity on a common logit scale. Model-data fit was evaluated using infit and outfit mean square (MNSQ) statistics. The results indicated adequate model fit and showed that most raters demonstrated acceptable scoring consistency. Rater severity estimates were closely clustered around the mean, indicating minimal evidence of systematic severity or leniency effects. Overall, the findings suggest that the scoring of the 2024 WASSCE Mathematics essay items was largely consistent, with limited rater-induced measurement error. The study highlights the usefulness of the MFRM model for monitoring rater behaviour and enhancing fairness in high-stakes examination scoring.

Keywords: Many-Facet Rasch Measurement; rater consistency; rater severity; rater leniency; WASSCE Mathematics

Introduction

Assessment outcomes play a critical role in educational decision-making, particularly in high-stakes examinations where results determine students' access to further education and career opportunities. In many educational systems, including those within West Africa, essay-type items are widely used in Mathematics examinations to assess higher-order cognitive skills such as reasoning, problem-solving, and logical justification. Despite their pedagogical value, essay items rely heavily on subjective human judgment during scoring, thereby increasing susceptibility to measurement error (Brookhart, 2018; Engelhard & Wind, 2018).

Rater-mediated assessment introduces potential inconsistencies arising from raters' interpretation of scoring rubrics, personal standards, fatigue, and prior expectations of examinee performance. These inconsistencies often manifest as rater severity or leniency, whereby some raters systematically assign lower or higher scores than others for comparable performances. When such rater effects are not adequately controlled, they threaten the reliability, validity, and fairness of examination scores (Eckes, 2015; Wind & Peterson, 2018). This concern is particularly pronounced in large-scale public examinations such as the West African Senior School Certificate Examination (WASSCE), which assesses a large population of candidates annually.

The WASSCE Mathematics examination comprises both objective and essay components, with the essay section designed to capture students' depth of understanding and procedural competence. Scoring of these responses is undertaken by trained Mathematics teachers using standardized marking guides provided by the West African Examinations Council (WAEC). Although marking schemes and rater training constitute essential quality-assurance measures, empirical evidence suggests that these procedures alone may not completely eliminate rater-induced measurement error (Engelhard & Wind, 2018; Eckes, 2015). Consequently, there is a need for empirical evaluation of whether raters apply scoring criteria consistently and without systematic bias.

Traditional approaches to evaluating scoring reliability, largely rooted in Classical Test Theory (CTT), provide limited insight into the complexities of rater-mediated assessments. While indices such as inter-rater reliability coefficients offer summary measures of agreement, they do not disentangle the distinct contributions of examinee ability, item difficulty, and rater severity to observed scores. Furthermore, CTT-based indices are sample-dependent and lack measurement invariance across different raters or examinee groups (DeMars, 2018; Engelhard, 2013).

The Many-Facet Rasch Measurement (MFRM) model provides a robust alternative for addressing these limitations. As an extension of the Rasch model, MFRM allows for the simultaneous estimation of multiple facets—typically examinees, items, and raters on a common interval scale expressed in logits (Eckes, 2019; Linacre, 2021). By explicitly modelling rater severity and consistency, the MFRM model offers diagnostic information for evaluating scoring quality, identifying misfitting raters, and enhancing fairness in assessment outcomes. Although the MFRM model has been widely applied in language testing, medical education, and professional certification contexts, its application in large-scale secondary school Mathematics examinations in West Africa remains limited. Given the importance of WASSCE results for students' educational trajectories, empirical evidence on rater behaviour in this context is essential. This study therefore applied the Many-Facet Rasch Measurement model to examine rater consistency in the scoring of the 2024 WASSCE Mathematics essay items.

Rater-mediated assessment plays a critical role in evaluating complex learner outcomes, particularly in contexts where higher-order cognitive skills such as reasoning, problem-solving, and justification are assessed. Unlike objective testing formats, such assessments rely on human judgment, making them inherently susceptible to measurement error when observed scores deviate from examinees' true ability due to construct-irrelevant factors (Eckes, 2019; Wind & Peterson, 2018). Among the most widely documented sources of such error are rater severity and leniency, which reflect systematic tendencies of raters to assign lower or higher scores than warranted. These tendencies may stem from differences in raters' experience, interpretation of scoring rubrics, or internal standards, and can result in inequitable scoring outcomes when multiple raters are involved. In large-scale examinations, where double marking or statistical moderation may be limited, such variability poses a significant threat to the fairness and validity of assessment decisions.

In addition to severity and leniency, inconsistency in rater behaviour further undermines the reliability of assessment outcomes. Ratets may apply scoring criteria unevenly across scripts or over time, leading to unpredictable variations in assigned scores. This lack of consistency weakens confidence in the dependability of examination results and raises concerns about the defensibility of decisions based on such scores (Eckes, 2015; Wind & Peterson, 2018). Consequently, there is a growing emphasis in contemporary assessment research on the need to model and control rater effects rather than treating them as random error.

Traditional approaches to analysing scoring reliability, largely grounded in Classical Test Theory, provide limited insight into the complexities of rater-mediated assessments. While indices such as inter-rater reliability coefficients offer useful summaries of agreement, they do not disentangle the distinct contributions of examinee ability, item difficulty, and rater severity to observed scores. Moreover, such indices are sample-dependent and do not ensure

measurement invariance across different groups of raters or examinees (DeMars, 2018; Engelhard, 2013). These limitations have prompted increased adoption of modern measurement approaches that provide more detailed diagnostic information about the scoring process.

The Many-Facet Rasch Measurement (MFRM) model has emerged as a robust framework for analysing rater-mediated data by simultaneously estimating multiple facets of the assessment process. Extending the Rasch model, MFRM places examinees, items, and raters on a common interval scale, thereby enabling direct comparison of their relative contributions to observed scores (Eckes, 2019; Linacre, 1989). A key strength of the model lies in its capacity to estimate rater severity and identify misfitting raters whose scoring patterns deviate from model expectations. In addition, MFRM provides fit statistics that serve as diagnostic indicators of rating quality, supporting the monitoring and improvement of scoring processes in high-stakes contexts. Studies have demonstrated the effectiveness of MFRM in enhancing the validity and fairness of rater-mediated assessments across diverse domains, including language testing, professional certification, and educational assessment (Wind & Peterson, 2018; Engelhard & Wind, 2018). These studies consistently show that even trained raters may exhibit subtle but systematic differences in scoring behaviour, reinforcing the need for analytical approaches that explicitly model such effects. By identifying and, where necessary, adjusting for rater-related variability, MFRM contributes to improved score comparability and more equitable assessment outcomes. Despite its established utility, the application of MFRM in secondary school Mathematics examinations, particularly within the West African context, remains limited. Given the high-stakes nature of the West African Senior School Certificate Examination and its significance for students' educational progression, there is a clear need for empirical investigation of rater behaviour in this setting. The present study addresses this gap by applying the Many-Facet Rasch Measurement model to examine rater consistency, severity, and leniency in the scoring of WAEC Mathematics essay items, thereby extending the application of advanced measurement techniques to a context of considerable educational importance.

Essay-type Mathematics items are central to the WASSCE because they assess complex cognitive skills that cannot be adequately measured through objective tests alone. However, reliance on human judgment in scoring introduces rater-mediated measurement error, which poses a serious concern in a high-stakes examination where outcomes influence students' educational and career trajectories. Even with standardized marking schemes and rater training, differences in rater severity, leniency, and consistency may persist, potentially leading to systematic bias and threatening the reliability and validity of examination scores (Eckes, 2019; Wind & Peterson, 2018).

When rater effects are not empirically examined, it becomes difficult to determine whether observed score variations reflect true differences in candidate ability or

are attributable to rater-related influences. Most evaluations of scoring quality in large-scale West African examinations have relied on Classical Test Theory indices, which provide limited diagnostic information and cannot adequately disentangle examinee ability from rater severity (DeMars, 2018; Engelhard, 2013). Consequently, empirical evidence on rater behaviour in WASSCE Mathematics essay scoring remains limited. Although the Many-Facet Rasch Measurement (MFRM) model provides a rigorous framework for simultaneously estimating examinee ability, item difficulty, and rater severity (Eckes, 2019; Linacre, 1989), its application to WASSCE Mathematics essay scoring is scarce. This gap underscores the need for a systematic investigation of rater consistency in the scoring of the 2024 WASSCE Mathematics essay examination to enhance the fairness, reliability, and defensibility of its scoring outcomes.

The purpose of this study was to examine rater behaviour in the scoring of the 2024 West African Senior School Certificate Examination (WASSCE) Mathematics essay items using the Many-Facet Rasch Measurement (MFRM) model. Specifically, the study sought to evaluate the consistency of raters, as well as the extent to which rater severity and leniency influence the scoring process. By modelling examinee ability, item difficulty, and rater effects simultaneously, the study aimed to provide empirical evidence on the fairness and reliability of essay scoring in a high-stakes examination context.

Research Questions

1. Are raters consistent when assigning scores in the 2024 WASSCE Mathematics essay items?
2. To what extent does rater severity occur in the scoring of the 2024 WASSCE Mathematics essay items?
3. To what extent does rater leniency occur in the scoring of the 2024 WASSCE Mathematics essay items?

Methodology

This study adopted a quantitative ex post facto design to examine rater behaviour in the scoring of the 2024 West African Senior School Certificate Examination (WASSCE) Mathematics essay items using the Many-Facet Rasch Measurement (MFRM) model. The design was appropriate as no variables were manipulated; rather, students' responses were obtained under standard examination conditions and subsequently scored independently by raters using established marking guidelines. The study utilised 75 Senior Secondary School III students' essay scripts, selected through purposive sampling to ensure completeness and suitability for analysis. Fifteen Mathematics teachers were selected through stratified random sampling from public secondary schools in Alimosho Local Government Area of Lagos State, Nigeria, with stratification based on teaching experience. All raters had prior experience in teaching and assessing senior secondary Mathematics and independently scored all scripts, resulting in a fully crossed rater-by-examinee design appropriate for MFRM analysis.

The instrument comprised the 2024 WASSCE Mathematics essay items and an accompanying scoring rubric aligned with WAEC marking guidelines. Content validity of the rubric was established through expert review by three specialists in Mathematics Education and Measurement and Evaluation. Item-level and scale-level Content Validity Indices (I-CVI and S-CVI) were computed, yielding an average S-CVI of 0.87, indicating acceptable content validity. Reliability of the scoring process was examined within the MFRM framework using rater reliability and separation indices, which indicated a high level of scoring consistency among raters. Data were analysed using FACETS® Rasch Measurement software (version 3.84.2). The model included examinees, items, and raters as facets, with estimates reported on a common logit scale. Model–data fit was evaluated using infit and outfit mean square statistics, with acceptable values ranging from 0.50 to 1.50. Rater consistency, severity, and leniency were examined using fit statistics, separation indices, and severity measures relative to the grand mean.

Results

Research Question One: Are raters consistent when assigning scores in the 2024 WASSCE Mathematics essay items?

Rater consistency was examined using infit and outfit mean square (MNSQ) statistics generated from the Many-Facet Rasch Measurement (MFRM) analysis. Acceptable fit was defined within the range of 0.50 to 1.50.

Table 1: Fit Statistics that show the consistency of raters when assigning ratings in 2024 WASSCE Mathematics Essay Items.

Rater (R)	Severity Measures (Logits)	Model Standard Error	Infit Mean-Square	Outfit Square	Mean-Square
5	1.54	0.12	0.53	0.51	
11	0.98	0.12	1.04	1.02	
6	0.93	0.13	0.97	0.94	
2	0.67	0.13	1.67	1.52	
8	0.60	0.13	1.06	1.05	
13	0.43	0.13	1.46	1.37	
15	0.32	0.13	0.95	0.94	
10	0.21	0.13	0.77	0.78	
7	-0.11	0.13	0.69	0.71	
4	-0.15	0.13	0.95	0.94	
9	-0.40	0.14	1.08	1.13	
14	-0.49	0.14	1.23	1.28	
3	-0.58	0.14	0.91	0.74	
12	-0.92	0.15	0.98	1.29	
1	-2.38	0.17	0.84	0.75	

Table 1 presents the fit statistics for the fifteen raters. The results show that fourteen out of the fifteen raters demonstrated acceptable infit and outfit values within the recommended range, indicating consistent application of the scoring criteria.

However, Rater 2 exhibited misfit, with an infit mean square value of 1.67 and outfit mean square value of 1.52, suggesting some inconsistency in scoring behaviour. Overall, the average infit and outfit mean square values of 1.01 and 0.99 respectively indicate good model fit and confirm that raters were generally consistent in assigning scores.

Research Question Two: To what extent does rater severity occur in the scoring of the 2024 WASSCE Mathematics essay items?

Rater severity was examined using the severity measures (logits) estimated from the MFRM model. Positive logit values indicate more severe raters, while negative values indicate more lenient raters.

Table 2: Rater Severity and Leniency Measures in Scoring 2024 WASSCE Mathematics Essay Items

Rater (R)	Severity Measures (Logits)	Model Standard Error	Infit Mean-Square	Outfit Mean-Square	Fair Average	Observed Average	Total Count
5	1.54	0.12	0.53	0.51	2.62	2.58	100
11	0.98	0.12	1.04	1.02	2.85	2.90	100
6	0.93	0.13	0.97	0.94	2.87	2.94	100
2	0.67	0.13	1.67	1.62	3.00	2.87	100
8	0.60	0.13	1.06	1.05	3.09	3.08	100
13	0.43	0.13	1.46	1.37	3.15	3.20	100
15	0.32	0.13	0.95	0.94	3.28	3.24	100
10	0.21	0.13	0.77	0.78	3.30	3.28	100
7	-0.11	0.13	0.69	0.71	3.37	3.34	100
4	-0.15	0.13	0.95	0.94	3.40	3.36	100
9	-0.40	0.14	1.08	1.13	3.54	3.54	100
14	-0.49	0.14	1.23	1.28	3.82	3.74	100
3	-0.58	0.14	0.91	0.74	3.97	3.90	100
12	-0.92	0.15	0.98	1.29	4.33	4.22	100
1	-2.38	0.17	0.84	0.75	4.65	4.52	100

As shown in Table 2, rater severity estimates ranged from 1.54 logits (most severe) to -2.38 logits (most lenient). Despite this range, most raters were clustered around the grand mean of 0.04 logits, indicating that differences in severity were relatively small.

Rater 5 was identified as the most severe, consistently awarding lower scores compared to other raters, while Rater 1 was the most lenient, awarding relatively higher scores. However, the proximity of most raters' severity estimates to the mean suggests that the overall level of severity was low and did not substantially distort scoring outcomes.

Research Question Three: To what extent does rater leniency occur in the scoring of the 2024 WASSCE Mathematics essay items?

Rater leniency was further examined by comparing observed average scores with fair average scores in the MFRM output.

As presented in Table 2, the observed averages and fair averages for most raters were closely aligned, indicating minimal deviation between expected and actual scoring behaviour. For instance, raters such as Rater 8 and Rater 10 showed nearly identical observed and fair averages, reflecting consistent and unbiased scoring. Although slight differences were observed among some raters, these variations were not substantial enough to indicate systematic leniency.

Overall, the findings suggest that rater leniency was minimal, and raters generally adhered closely to the expected scoring standards.

Discussion of Findings

This study examined rater consistency, severity, and leniency in the scoring of the 2024 WASSCE Mathematics essay items using the Many-Facet Rasch Measurement (MFRM) model. The findings provide empirical evidence that raters generally applied the scoring criteria consistently, with minimal rater-induced measurement error. Most raters demonstrated acceptable fit to the MFRM model, indicating stable and predictable scoring behaviour across examinees and items. This finding supports the reliability and fairness of the essay scoring process in a high-stakes examination context and aligns with recent studies emphasizing the importance of monitoring rater effects in performance-based assessments (Wind & Peterson, 2018; Eckes, 2015).

The infit and outfit mean square statistics revealed that fourteen of the fifteen raters fell within the recommended fit range of 0.50 to 1.50, with only one rater exhibiting misfit. Misfitting raters are often indicative of inconsistencies in the application of scoring criteria or deviations from expected rating patterns (Linacre, 2021; Engelhard & Wind, 2018). However, the presence of a single misfitting rater does not substantially compromise overall scoring quality. Rather, it highlights the diagnostic strength of the MFRM model in identifying raters who may require additional training or calibration. Similar findings have been reported in recent empirical studies, which show that isolated instances of rater misfit are common but typically do not undermine the overall integrity of scoring systems when the majority of raters demonstrate acceptable fit (Eckes, 2019; Engelhard & Wind, 2018).

Rater severity and leniency estimates were closely clustered around the grand mean, indicating that raters operated within a shared scoring standard. Although some variation in severity was observed, the magnitude of these differences was relatively small and unlikely to have a substantial impact on examinee scores. This suggests that the standardized marking guides and rater training procedures were effective in minimizing systematic differences in scoring behaviour. Recent research supports this finding, indicating that structured rubrics and systematic rater training contribute significantly to reducing variability in rater severity and

improving scoring consistency in large-scale assessments (Wind & Peterson, 2018; Engelhard, 2018).

The low level of rater-induced measurement error observed in this study has important implications for fairness in high-stakes examinations. When rater effects are minimal, observed scores more accurately reflect true differences in examinee ability rather than scorer-related influences. This is particularly critical in the WASSCE context, where examination outcomes play a decisive role in students' educational progression and future opportunities. The findings therefore provide empirical support for the defensibility and credibility of the scoring process used in the 2024 WASSCE Mathematics essay examination.

Methodologically, the study reinforces the usefulness of the MFRM model as a robust framework for evaluating rater-mediated assessments. Unlike Classical Test Theory approaches, which provide only global indices of agreement, MFRM offers detailed diagnostic information on individual raters while simultaneously modelling examinee ability, item difficulty, and rater severity on a common measurement scale (Eckes, 2019; Linacre, 2021). This enhances the interpretability, transparency, and fairness of assessment outcomes, making it particularly suitable for high-stakes examination contexts.

Conclusion

This study demonstrated the effectiveness of the Many-Facet Rasch Measurement (MFRM) model in evaluating rater behaviour in a high-stakes Mathematics essay examination. The findings indicate that the scoring of the 2024 WASSCE Mathematics essay items was characterized by a high level of rater consistency and minimal rater-induced measurement error. The majority of raters exhibited acceptable model fit, and rater severity estimates were closely clustered around the mean, suggesting limited evidence of systematic severity or leniency effects. These results imply that examinee scores largely reflected true differences in mathematical ability rather than scorer-related bias, thereby supporting the fairness and reliability of the scoring process. The study contributes to the growing body of evidence supporting the application of Rasch-based approaches in monitoring and improving rater-mediated assessments. Importantly, the findings underscore the value of integrating MFRM into routine quality assurance practices in large-scale examinations. By providing detailed diagnostic information on rater behaviour, the model offers examination bodies a robust framework for enhancing scoring consistency, identifying potential sources of bias, and strengthening the credibility and defensibility of assessment outcomes.

Recommendations

Examination bodies such as WAEC are encouraged to institutionalize Rasch-based monitoring of rater behaviour as part of routine quality-assurance practices. Continuous rater training supported by statistical diagnostics is recommended to sustain fair, reliable, and defensible essay scoring in high-stakes examinations.

References

- Brookhart, S. M. (2018). *How to create and use rubrics for formative assessment and grading*. ASCD.
- DeMars, C. (2018). *Item response theory* (2nd ed.). Oxford University Press.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.
- Eckes, T. (2019). Many-facet Rasch measurement. In J. I. Liontas (Ed.), *The Routledge handbook of language testing*. Routledge.
- Engelhard, G. Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Engelhard, G. Jr., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.
- Linacre, J. M. (2021). *FACETS Rasch measurement computer program user's guide* (Version 3.84.2).
- Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rater effects. *Educational Measurement: Issues and Practice*, 37(3), 28–47.
- West African Examinations Council (WAEC). (2024). *Chief examiners' report: West African Senior School Certificate Examination (WASSCE) Mathematics*. WAEC.