Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Application of explanatory Rasch model: Standardizing West-African Certificate Examination

D. O. Oyeniran [*†], J. Dadzie [†], C. Ocheni [†], and Y. Isah [†]
†The University of Alabama, Tuscaloosa, Alabama, United States
*Corresponding author. Email: dooyeniran@crimson.ua.edu

Abstract

This study applies the Dichotomous Many-Facet Rasch Model (D-MFRM) to evaluate the fairness and psychometric quality of the Basic Education Certificate Examination (BECE) mathematics test in a West African context. Using a balanced sample of 1,168 students across gender and school type, the analysis examined item difficulty, person ability, and potential differential item functioning (DIF). Results indicate strong model fit, with Infit and Outfit Mean Square Errors near 1.0 and standardized residuals within ±2.0. Items were well distributed across the ability spectrum, targeting average and lower-ability students. DIF analyses revealed minimal bias across gender and school type, confirming measurement invariance. While item reliability was high (0.97), student separation reliability was low, suggesting limited differentiation among individual abilities. These findings validate the BECE mathematics test as an equitable assessment tool and highlight D-MFRM's utility for ensuring fairness in large-scale educational testing.

Keywords: Exploratory Rasch Model, Many Facet Rasch Model, West-African Examination

Large-scale educational assessments hold transformative potential within national education systems, serving multiple crucial roles. They help policymakers monitor student performance, evaluate educational equity, and guide resource distribution (Lietz & Tobin, 2016). Tests such as Basic Educational Certificate Examination (BECE), Programme of International Students Achievement (PISA), and Trend in International Mathematics and Science Study (TIMSS) offer standardized benchmarks to ensure comparability across diverse schools and regions (Azim Premji University, 2024). These assessments are high stakes, influencing student promotion, placement, teacher accountability, and school funding (Loeb & Byun, 2019).

Test standardization ensures validity and fairness by reducing bias through consistent item development, psychometric calibration, and measurement invariance. It helps ensures score differences reflect real ability, not irrelevant constructs (AERA, APA & NCME, 2014). Fairness procedures including DIF detection, support equitable performance across groups and underpin high stakes decisions in education, policy alignment, and resource allocation in large-scale assessments (Dorans & Cook, 2016).

The Many-Facet Rasch Model (MFRM) extends the Rasch framework by incorporating additional facets that influence test responses, such as item features, testing conditions, and contextual factors. The Dichotomous MFRM (D-MFRM) is particularly useful for detecting DIF, where items favor certain groups due to factors unrelated to ability. Previous studies document gender differences and school-type disparities linked to resources and instruction (Kim & Wilson, 2020; Mohamat et al., 2022; Syafrial et al., 2022). This study applies the D-MFRM to control bias and rigorously evaluate item and person fit. The D-MFRM expresses the log-odds of a correct response as a linear combination of parameters, as shown in equation 1 below.

$$\ln\left[\frac{P_{nim(x=1)}}{P_{nim(x=0)}}\right] = \theta_n - \delta_i - \eta_m, \quad (1)$$

Where P_{nim} is the probability that person n succeeds on item i under facet conditions m, $\theta_n$is the ability of person n, δ_i is the difficulty of item i while $\eta_m$is the effects of other facets (e.g., gender, school type). This model ensures measurement invariance by adjusting for systematic influences from multiple facets, making it ideal for fairness analysis in large-scale assessments.

Thus, this study uses the MFRM to ensure psychometrically fair, invariant, and comparable high-stakes assessments across gender and school type, supporting equity and trust in BECE outcomes in West Africa.

1. Method

1. 1 Data Source and Samples

The study utilized responses from Basic Education Certificate Examination (BECE) mathematics items administered to students in a large-scale testing program within a West-African country. Although the original BECE administration in 2022 involved over 550,000 responses, a randomly sampled subset was used for diagnostic analysis. This subset included 40 multiple-choice items and ensured a balanced representation of gender with 598 males and 570 females, and based on school type there are 584 students each from public and private schools.

1. 2 Model Application

A MFRM was used to jointly estimate student ability, item difficulty, and the influence of gender and school type. The model was calibrated on a logit scale, allowing for evaluation of person fit to identify inattentive or erratic responses, and to determine if item fit conforms to Rasch assumptions. We conducted group comparisons between males and females, and between public and private schools, to detect DIF attributable to demographic factors rather than true differences in mathematical ability.

1. 3 Analytical Techniques

The analysis employed several diagnostic tools. Wright Maps were used to display the distribution of person abilities alongside item difficulties. Boxplots of Infit and Outfit Mean Square Errors (MSE) provided visual confirmation of the data-to-model fit, with values around 1.0 indicating acceptable fit. Standardized residual plots helped to identify any outlier items or group discrepancies by showing residuals centered around zero. Calibration tables summarized the logit-scale estimates, standard errors, and separation reliability indices for the different facets involved.

2. Result

The results from the dichotomous Rasch model indicated that the items were generally well distributed across the ability spectrum. Approximately 20 of the 40 items were identified as “easy” across all groups. The overall Infit and Outfit MSE values were near the ideal of 1.0, and standardized residuals remained within the acceptable range of ±2.0. These findings suggest that the model demonstrated a good fit with the observed data.

Table 1. Summary result

Statistics	Students	School type	Class	Items
Logit Scale Location Mean	0.38	0.00	0.00	0.00
Logit Scale Location SD	0.32	0.01	0.01	0.38

Standard Error Mean	0.33	0.01	0.01	0.05
Standard Error SD	0.01	0.00	0.00	0.04
Outfit MSE Mean	1.00	1.07	1.07	1.07
Outfit MSE SD	0.07	0.06	0.06	0.06
Infit MSE Mean	1.00	1.07	1.07	1.07
Infit MSE SD	0.06	0.06	0.06	0.06
Std. Outfit Mean	0.02	1.65	1.65	1.65
Std. Outfit SD	0.80	1.15	1.15	1.15
Std. Infit Mean	0.02	1.65	1.65	1.65
Std. Infit SD	0.77	1.15	1.15	1.15
Separation Reliability	-0.06	0.48	-1.30	0.97

Calibration data in table 1 showed that students had a higher average location on the logit scale (M = 0.38, SD = 0.32) compared to school type, class, and items, which were centered near zero. This indicates that the student abilities were distinctly measured relative to the baseline established by the items. However, while item separation reliability was high (0.97), the separation reliability for students was poor (−0.06), implying that the model had difficulty differentiating between individual student abilities. Similar issues were noted for class reliability, whereas school type reliability (0.48) suggested moderate differences among school types.

Table 2. Subgroup fit

Groups	Average Rating	Subgroup Location	Subgroup Location SE	Outfit MSE	Std. Outfit	Infit MSE	Std. Infit
public	0.59	0.01	0.01	0.99	-0.02	1.00	-0.01
private	0.59	-0.01	0.01	1.00	0.05	1.00	0.05
male	0.59	0.00	0.01	1.00	0.01	1.00	0.01
female	0.59	0.00	0.01	1.00	0.02	1.00	0.02

Table 2 show the relative ordering and locations of the groups on the logit scale in terms of difficulty. The results indicate that all the subgroups (public, private, male, and female) have similar difficulty (average rating = 0.59). Additionally, the result shows that the private school has the lowest location (δ =-0.01) on the logit scale while the public school has the highest location (δ =0.01), whereas, males and females share the same location (δ =0.0). This result shows that there is relatively no difference among the groups in terms of their location, implying that groups have somewhat the same location on the logit scale. The Infit and Outfit *MSE* for all the groups are approximately at 1.00, an indication of acceptable fit statistics values for the various groups. These results suggest that there is a good model to fit data and consistency in response patterns across the groups. An indication that the scale is functioning appropriately for all the various subgroups with little to no measurement error.

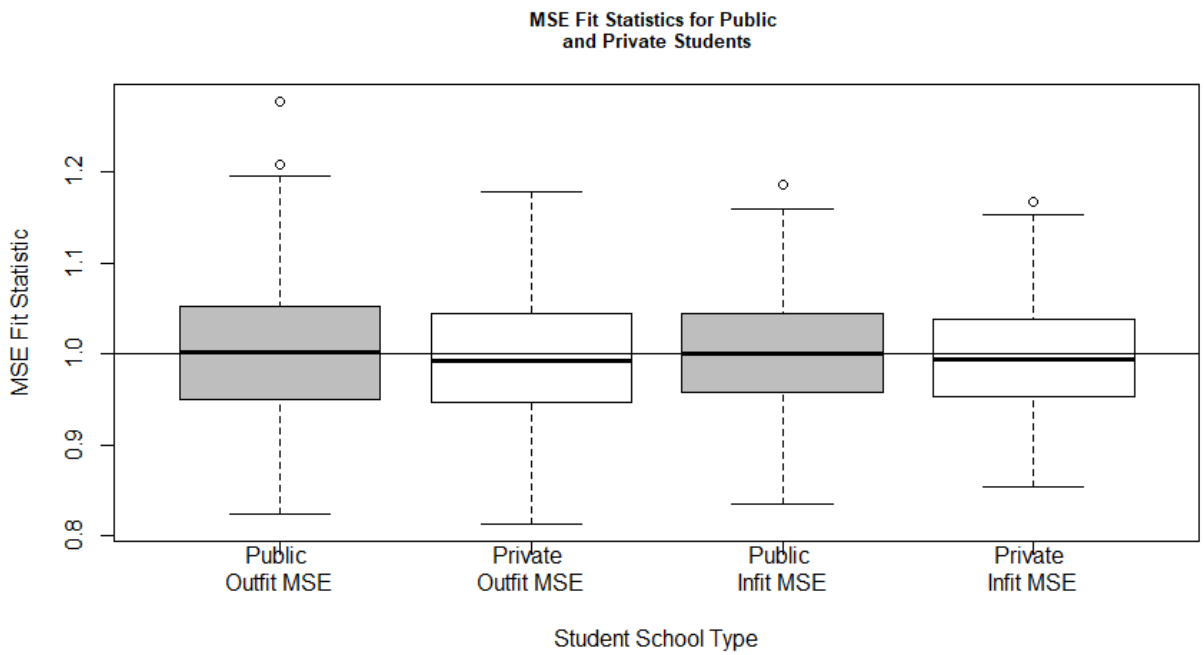

Figure 1. Student school type -MSE

Figure 1 shows the boxplot of the MSE fit statistics for public and private school students. The plot shows that the MSE fit statistics are centered around 1.0, which is the expected value for good model fit in Rasch measurement. A comparison of these fit statistics between private (white box) and public (grey box) school students shows quite similar distributions in both Infit and Outfit MSE. These results suggest a consistent and similar response pattern among students in both private and public schools. Consequently, the box plot in Figure 2 shows standardized residuals (*Zstd*) which are centered around 0.00 for both private (white box) and public (grey box) school students, an indication that the responses align well with the model's expected responses. This suggests that there is no significant difference between public and private school students in terms of the fit indices, implying that there are similar response behaviors between students from both private and public schools.

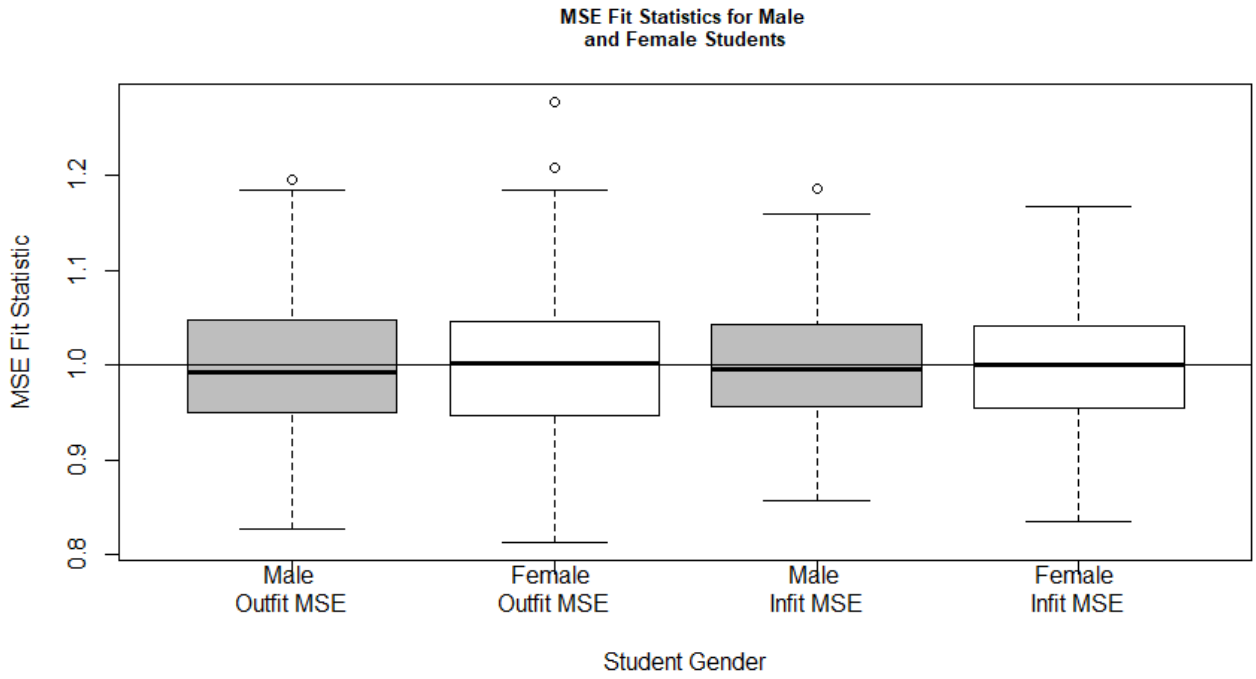

Figure 2. Student Gender -MSE

The boxplot in Figure 2 shows the MSE fit statistics for male and female students. The MSE fit values centered around 1.00, suggest acceptable fit values. When these fit statistics are compared between male (grey box) and female (white box) students, it suggests similar distributions in both Infit and Outfit MSE. This result is an indication of consistent response patterns between male and female students. Furthermore, the box plot in Figure 4 shows standardized residuals (*Zstd*) which are centered around 0.00 for male (grey box) and female (white box) students, an indication that the responses align well with the model's expected responses. This suggests that although there are noticeable outliers, there is no significant difference between male and female students in terms of the fit indices, implying that there are similar response patterns between male and female students.

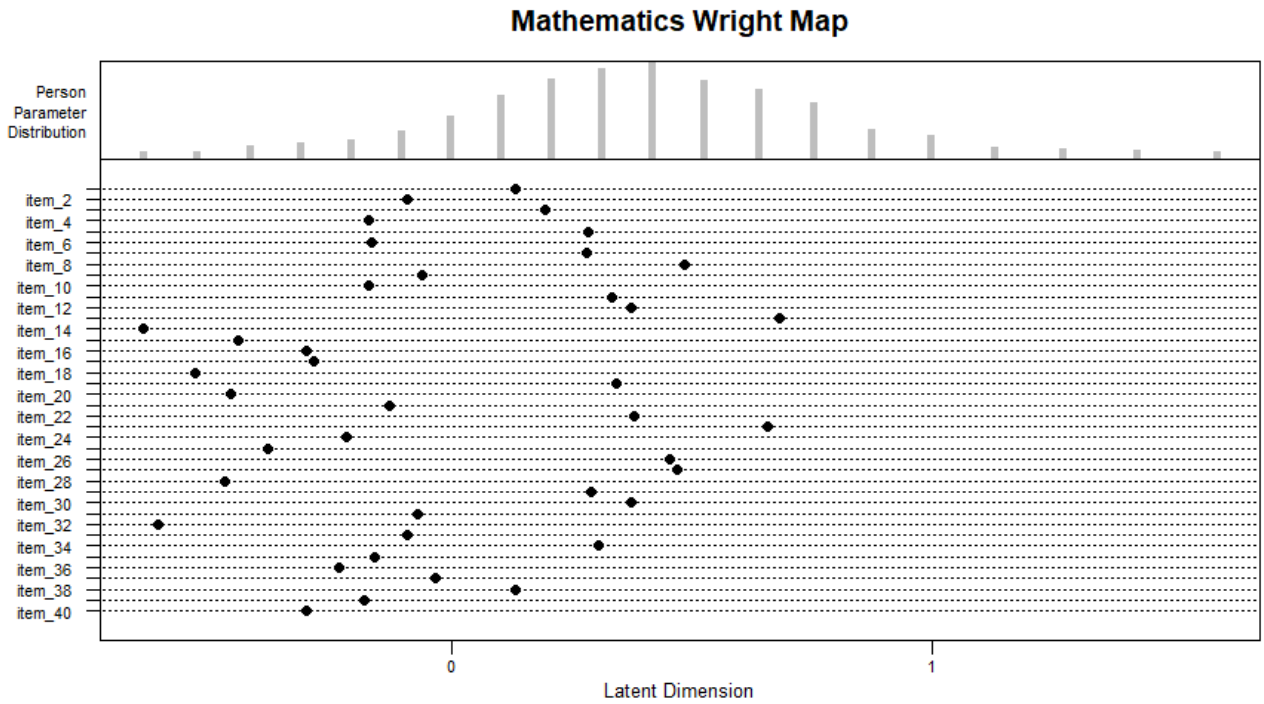

Figure 3. Wright Map All items

Figure 3 presents the person-item (Wright) map for the mathematics examination, displaying the distribution of students' abilities and item difficulties along a common latent scale. The histogram shows students' mathematical ability, with higher bars indicating higher ability and lower bars indicating lower ability. Item locations, shown as black dots, reflect their difficulty levels, with harder items positioned to the right and easier items to the left. The results indicate that students' abilities are largely centered around zero, suggesting an average ability level for students. However, the observed gap between item difficulty and person ability implies that the test may not be optimally targeted to the population.

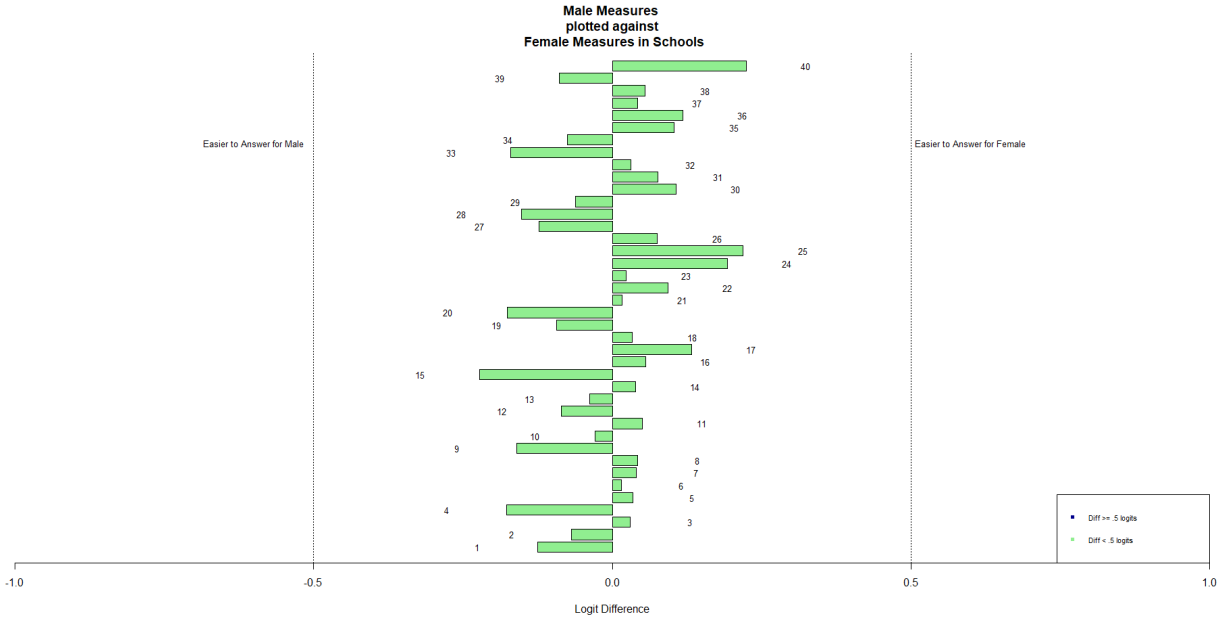

Figure 4. Gender Group comparison

To further substantiate the results, the bar plot in Figure 4 shows no evidence of DIF by gender, as all items fall within the ±0.50 DIF threshold, indicating comparable item performance for males and females.

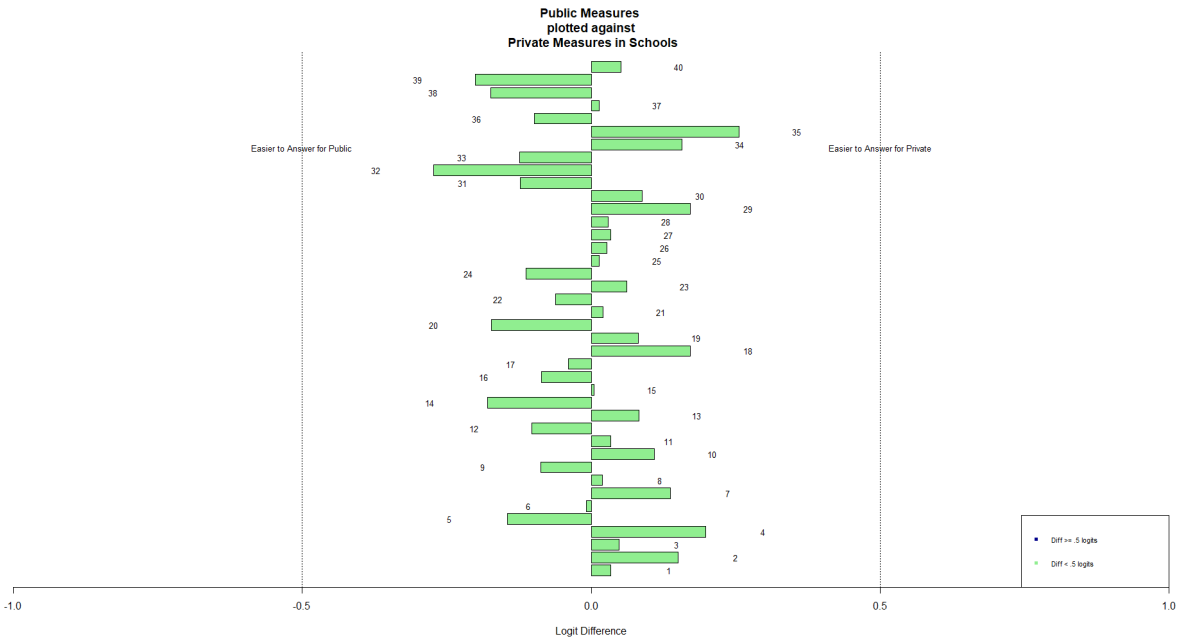

Figure 5. School Type Comparison

The bar plot in figure 5 further shows that there is no evidence to support DIF for the groups as the green bars are all within the boundaries of -0.50 and +0.50 set as regions of substantial DIF for public and private schools. In essence, some items appear easier for public schools while others are easier for private schools, but there is no sufficient evidence to suggest any significant difference between public and private schools.

3. Discussion

The findings from this study underscore the efficacy of the D-MFRM in handling complex measurement challenges in educational assessment. By incorporating multiple facets beyond just person and item parameters the model provides a nuanced understanding of how various external factors (such as gender and school type) interact with test performance. Despite previous concerns regarding potential biases in large-scale assessments, the results disagreed with the findings of Kim and Wilson (2020) and Mohamat et al. (2022), indicating that the BECE mathematics test items exhibit no substantial differential item functioning across gender and school type.

The model's ability to capture the subtle interplay between person ability and item difficulty is particularly noteworthy. Although some minor deviations were observed in a few items, these were not consistent enough to suggest any systematic bias. This reinforces the importance of using advanced psychometric models like the D-MFRM, which can pinpoint issues that might be overlooked by traditional statistical methods. The study also highlights the critical role of person and item fit analyses; ensuring that both individual responses and test items conform to the model expectations is essential for maintaining the validity and reliability of test scores.

Furthermore, while private schools are often presumed to provide superior resources leading to better educational outcomes, the data from this study reveal that the measured mathematical abilities of students across public and private schools are remarkably similar, disagreeing with the reports of Kim and Wilson (2020), and Syafrial et al. (2022). This finding suggests that factors beyond the immediate test environment, such as broader socio-economic variables and instructional quality, might play a more decisive role in shaping educational outcomes.

4. Conclusion

In summary, the application of the D-MFRM to the BECE mathematics test data demonstrates that the model is a robust tool for analyzing dichotomous responses in educational assessments. The study confirmed that the test items are well calibrated and exhibit good overall model fit. There is no substantial evidence of differential item functioning based on gender or school type. Both public and private school students show similar distributions of mathematical ability, with items targeting average and lower-ability ranges.

These conclusions validate the use of the D-MFRM in high-stakes testing environments, offering educators and researchers a powerful method for ensuring measurement invariance across diverse groups.

4. Implication

This study has important implications for educational practice and policy. The lack of substantial DIF indicates that the BECE mathematics test functions as a fair and unbiased measure across demographic groups, supporting its use for high-stakes decisions. Policymakers can therefore direct attention toward other sources of inequality, such as resource distribution, instructional quality, and socio-economic conditions. Methodologically, the study highlights the value of the D-MFRM for examining assessment fairness and encourages its use in similar contexts. Future research may incorporate additional facets to capture broader testing influences. For educators, the results affirm that instructional and curricular decisions are based on equitable assessment outcomes.

Acknowledgement

Funding Statement: The authors did not receive funding to support this research.

Competing Interests: The authors have no conflict of interest.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Azim Premji University. (2024). *Using large-scale assessments to inform education policy and governance*.

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in Educational Assessment and Measurement*. Routledge.

Kim, J. & Wilson, M. (2020). Polytomous item explanatory item response theory model. *Educational and Psychological Measurement, 80*(4), 726-755. https://doi.org/10.1177/0013164419892667

Lietz, P., & Tobin, M. (2016). *The impact of large-scale assessments in education on education policy: evidence from around the world*. *Research Papers in Education, 31*(5), 499–501. https://doi.org/10.1080/02671522.2016.1225918

Loeb, S., & Byun, E. (2019). *Testing, Accountability, and School Improvement*. *Annals of the American Academy of Political and Social Science, 683*(1), 94–116. https://www.jstor.org/stable/26732341

Mohamat, R., Sumintono, B., & Abd Hamid, H. S. (2022). Raters' Assessment Quality in Measuring Teachers' Competency in Classroom Assessment: Application of Many Facet Rasch Model. *Asian Journal of Assessment in Teaching and Learning*, *12*(2), 71–88. https://doi.org/10.37134/ajatel.vol12.2.7.2022

Simon, M., Ercikan, K., & Rousseau, M. (2013). *Improving Large-Scale Assessment in Education: Theory, Issues, and Practice*. Routledge.

Syafrial., Ashadi., Saputro, S., & Sarwanto. (2022). Trend creative thinking perception of students in learning natural science: Gender and domicile perspective. *International Journal of Instruction*, *15*(1), 701-716. https://doi.org/10.29333/iji.2022.15140a