



Rasch Analysis of Item Fit and Differential Functioning of Senior Secondary School Mathematics Test in Rivers State

EJIMAJI, F. U.¹; OBILOR, E. I.² & WAGBARA, E. S.³

^{1,2,3} Department of Educational Foundations, Faculty of Education, Rivers State University, Port Harcourt

Article history:

Received: 05/07/2026

Accepted: 10/08/2026

Published: 03/10/2026

Keywords: Rasch model, analysis, item fit, differential item functioning.

***Corresponding Author:**

EJIMAJI, F. U.

Abstract

This study analysed item fit and differential functioning of senior secondary school mathematics test in Rivers State using the Rasch model. Guided by two research questions and null hypotheses, the study adopted an instrumentation research design. The population consisted of 199,602 Rivers State public senior secondary school students for the 2025/2026 academic year, from which a sample of 480 participants was drawn using a multi-stage sampling technique. A researcher- developed instrument titled “Mathematics Performance Test for Senior Secondary School Two Students”, was used for data collection. The instrument was validated by two senior secondary school mathematics teachers and a Measurement and Evaluation expert at Rivers State University. The reliability coefficient was calculated using Kuder–Richardson Formula 20, yielding $\alpha = 0.73$. Data were analysed using the Rasch model with the eRm package in R. Results showed that all the items of the senior secondary school mathematics test demonstrated acceptable fit to the Rasch measurement model, and the test items do not demonstrate differential item functioning by gender. It was recommended that mathematics educators collaborate with measurement specialists to utilise the Rasch model throughout item construction to detect and evaluate possible item misfit and guarantee that all items exhibit adequate fit. Also, schools should always review their mathematics test items for differential item functioning by gender to be satisfied that the test does not introduce gender-related bias before administering to students.

Original Research Article

Copyright © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0)

How to cite this article: EJIMAJI, F. U., OBILOR, E. I., & WAGBARA, E. S. (2026). Rasch analysis of item fit and differential functioning of senior secondary school mathematics test in Rivers State. EIRA Journal of Arts, Law and Educational Sciences, 2(5), 18–27.

Introduction

In every career, basic knowledge of mathematics is non-negotiable, as it equips individuals with the analytical, problem-solving, and logical skills and reasoning to navigate career and life challenges. Mathematics is the foundation upon which scientific and technological innovations are built, and it is a prerequisite for admission into many tertiary education programmes, especially in science-based disciplines such as engineering, medicine, computer science, economics, architecture, and environmental sciences. The importance of mathematics in the Nigerian educational system is evident in the national educational policies and curriculum frameworks, which emphasise the need for students to acquire quantitative reasoning, logical thinking, and analytical skills necessary for national growth and development (Federal Republic of Nigeria, 2014). Despite the critical role of the subject in human existence, it is found that many secondary school students perform poorly in examinations such as the Senior School Certificate Examination organised by the National Examinations

Council (NECO) and the West African School Certificate Examination organised by the West African Examinations Council (WAEC) whenever the examination results are released, as evident in the reports of the 2023 WAEC and 2025 NECO results, which showed low performance among secondary school students in mathematics. Different factors, including inadequate instructional methods, lack of qualified teachers, poor learning environments, students' anxiety toward mathematics, inadequate instructional materials, and weaknesses in assessment practices, have been identified as possible reasons for poor performance of students in mathematics (Adedoyin & Soyibo, 2021). Obviously, among these factors, educational assessment seems to remain the most important factor because it directly influences teaching, learning, curriculum implementation, educational policy, and decision-making processes. It is also held that without comprehensive psychometric evaluation, it becomes difficult to determine whether mathematics tests accurately measure students' mathematical abilities (Akpan & Obilor, 2021);

therefore, there is a need to examine the fitness and differential functioning of mathematics test items to ascertain whether or not the low performance of students in the subject is due to the quality of the test item in terms of item fit and differential item functioning (DIF).

Item fit is an important quality of Rasch model measurement. Item fit refers to how well each test item aligns with the expectations of the Rasch model, which indicates whether students' responses to an item follow a logical pattern consistent with their overall ability or not. Fit statistics allow educators to refine tests, remove biased items, and detect anomalies such as guessing or disengaged students (Retnawati et al., 2021). Misfitting items in Nigerian mathematics exams have been linked to ambiguous phrasing or culturally unfamiliar contexts (Ojerinde, 2020). Item fit is a central concept in modern educational measurement, particularly within Item Response Theory (IRT) and Rasch modelling frameworks. Item fit is defined as the degree of agreement between observed student responses and responses predicted by a measurement model. It can also be described as the extent to which observed response patterns for each test item align with the expectations of a specified measurement model.

Rasch Fit Statistics and Item Diagnostics

Item fit and person fit statistics identify items or respondents that deviate from model expectations:

- a. Infit Mean Square (MNSQ): Sensitive to patterns near the person's ability level.
- b. Outfit MNSQ: Sensitive to unexpected extreme responses.
- c. Z-Standardized Fit Values: Used for hypothesis testing about misfit.

Fit statistics allow educators to refine tests, remove biased items, and detect anomalies such as guessing or disengaged students (Bond & Fox, 2015; Retnawati et al., 2021). Misfitting items in Nigerian mathematics exams have been linked to ambiguous phrasing or culturally unfamiliar contexts (Ojerinde, 2020).

Infit and Outfit Statistics

Within Rasch modelling, item fit is primarily evaluated using infit and outfit mean square (MNSQ) statistics as briefly discussed below:

- a. Infit Mean Square (Information-Weighted Fit): Infit is sensitive to unexpected responses from persons whose ability levels are close to the item difficulty level. Because these responses are most informative for measurement, infit is particularly useful for detecting subtle inconsistencies in item functioning (Bond & Fox, 2015). According to Ismail and Sumintono (2021), recent studies show that infit misfit in mathematics assessments often results from:

- i. unclear wording of word problems;
 - ii. multiple solution strategies; and
 - iii. cognitive ambiguity in item structure.
- b. Outfit Mean Square (Outlier-Sensitive Fit): Outfit is sensitive to extreme unexpected responses, such as guessing by low-ability students or careless mistakes by high-ability students (De Ayala, 2022). It has been suggested that outfit misfit is commonly associated with test anxiety and disengagement in low-stakes assessments (DeMars, 2020). In terms of the interpretations, Boone et al. (2019) and Linacre (2020) observed that while traditional thresholds (0.7–1.3 or 0.5–1.5) are widely used, modern psychometric research emphasises context-sensitive interpretation rather than rigid cut-offs.

Importance of Item Fit in Mathematics Assessment

Item fit analysis is particularly important in mathematics because mathematical ability is multidimensional, involving conceptual understanding, procedural fluency, and problem-solving skills (Kilpatrick et al., 2021). The key roles of item fit in mathematics assessment include:

- a. Cognitive Alignment: Item fit ensures that items measure intended cognitive processes rather than unrelated skills such as reading comprehension (Sullivan et al., 2020).
- b. Detection of Construct-Irrelevant Variance: Poorly fitting items often reflect external influences such as cultural context or linguistic complexity (Chan et al., 2020).
- c. Curriculum Alignment: Misfitting items may indicate gaps between curriculum objectives and assessment content (OECD, 2023).
- d. National Examination Quality: In systems such as WAEC and NECO, item fit analysis is essential for improving the validity and fairness of mathematics examinations (Ojerinde, 2020).

In Rasch analysis, item fit provides empirical evidence of whether items function consistently with the underlying latent trait being measured, such as mathematical ability (Bond & Fox, 2015). Misfitting items occur when response patterns deviate from this probabilistic expectation. Such deviations may arise due to multidimensionality, poor item construction, guessing, or construct-irrelevant variance (De Ayala, 2022). In mathematics education, item fit is particularly critical because test items vary in cognitive complexity, language demand, and problem-solving structure. When items do not conform to model expectations, they introduce measurement error and threaten the validity of score interpretations (Retnawati et al., 2021). Recent large-scale assessment literature also emphasises that item fit is essential for ensuring fairness, especially in high-stakes testing contexts such as national examinations (Zieky, 2021). Aminu and Bello

(2019) reported in their study that 42 items fit the model well, while three items exhibited infit and outfit mean square statistics above 1.5, suggesting misfit; hence, they recommended revision of misfitting items and continuous monitoring of item performance. Retnawati et al. (2021) observed that 10–15% of mathematics test items exhibited misfit due to ambiguity and cognitive overload. Linacre (2020) opined that while contemporary psychometric research shows that Rasch measurement supports fairness and measurement invariance in educational testing across demographic groups, other scholars argued that Rasch analysis provides more accurate estimates of students' mathematical abilities than Classical Test Theory because of its emphasis on invariant measurement and probabilistic modeling.

Differential Item Functioning (DIF) in Rasch Measurement and Educational Assessment

DIF occurs when students with the same underlying ability have different probabilities of success on an item due to group membership (e.g., gender, region, school type). Rasch analysis can identify both uniform DIF (consistent advantage) and non-uniform DIF (advantage varies across ability levels) (Zieky, 2021). In mathematics assessment, DIF can arise from cultural context, gendered phrasing, or embedded content that advantages specific groups. In educational assessment, DIF occurs when students with equivalent mathematical ability respond differently to a test item due to characteristics unrelated to the ability being measured, such as gender, socio-economic status, linguistic background, or school location (Hambleton et al., 2016). Integrating Rasch modelling with DIF detection would enable educators and researchers to identify and rectify biases arising from linguistic, cultural, gendered or instructional differences, ensuring that assessment scores reflect true mathematical ability rather than extraneous factors. This implies that rigorous test evaluation in mathematics requires a multi-faceted approach.

Differential item functioning (DIF) is more than just a statistical consideration; it embodies an ethical dimension in educational assessment. The DIF analysis is especially vital because students come from varied linguistic, socio-cultural, and instructional environments. Understanding DIF allows educational stakeholders to ensure that mathematics tests measure actual ability rather than contextual factors (Ojerinde, 2020). Differential item functioning occurs when two students with identical underlying abilities respond differently to a particular item because of group membership. Bond and Fox (2015) stated that DIF highlights violations of measurement invariance, a key principle in Rasch and IRT modelling. It is important to emphasise that DIF does not automatically indicate a defective item. Some differences may reflect legitimate group-based knowledge differences. However, in standardised mathematics testing, such as Senior Secondary School Certificate Examinations (SSSCE),

unexplained DIF threatens the validity of test scores and may produce inequitable outcomes (De Ayala, 2022).

Types of Differential Item Functioning

1. **Uniform Differential Item Functioning:** This occurs when one group consistently has higher or lower probabilities of answering an item correctly across all ability levels. For instance, urban students might consistently outperform rural students on word problems due to language familiarity, regardless of mathematical ability (Hambleton et al., 2016).
2. **Non-uniform Differential Item Functioning:** This arises when differences vary across ability levels. This may happen if one group performs better at low ability levels but worse at high ability levels, often due to interactions between cognitive strategies, item format, or contextual familiarity (Hidalgo et al., 2020).

Causes of Differential Item Functioning in Mathematics Tests

Research identifies several sources of DIF in mathematics assessment:

- a. **Language and Reading Load:** Even mathematically simple items can become challenging due to complex phrasing or unfamiliar vocabulary (Sullivan, Mousley, & Zevenbergen, 2020; Alderson et al., 2019).
- b. **Cultural Contexts:** Items referencing culturally specific experiences can disadvantage students from other backgrounds (Chan et al., 2020; Odili & Okeke, 2019).
- c. **Instructional Differences:** Variations in teaching quality and curriculum coverage create disparities in familiarity with certain item types (Hambleton et al., 2016).
- d. **Gender and Stereotype Influences:** Contexts perceived as “masculine” or “feminine” can influence problem-solving success (Zieky, 2021; Stoet & Geary, 2018).
- e. **Cognitive Strategy Variations:** Different learning experiences may favour certain problem-solving approaches (De Ayala, 2022; Penfield & Lam, 2018).

Methods of Detecting Differential Item Functioning

Various statistical approaches detect Differential Item Functioning: They include:

- a. **Rasch-Based Differential Item Functioning Analysis:** Compares item difficulty parameters across groups (Bond & Fox, 2015).
- b. **Mantel–Haenszel Method:** Classical approach using odds ratios to evaluate DIF (Hambleton et al., 2016).

- c. Logistic Regression: Examines interactions between ability and group membership (Hidalgo et al., 2020).
- d. Item Response Theory Likelihood Ratio Tests: Compares nested models to identify item-level DIF (De Ayala, 2022).
- e. Ordinal Logistic Approaches: Useful when items have partial credit scoring (Stark, Chernyshenko, & Drasgow, 2019).

Combining these methods often yields more robust DIF detection (OECD, 2023; Stoet & Geary, 2018).

Significance of Differential Item Functioning Analysis

The following are the significance of Differential Item Functioning:

1. Fairness: Ensures students' scores reflect ability rather than extraneous factors (Zieky, 2021).
2. Validity: Protects the integrity of test score interpretation (Penfield & Lam, 2018).
3. Policy and Accountability: Supports equitable educational policies and accurate placement decisions (Ojerinde, 2020).
4. Test Improvement: Guides revision or removal of biased items (Hambleton et al., 2016; Odili & Okeke, 2019).
5. International Comparability: DIF analysis underpins fairness in large-scale assessments like PISA or TIMSS (OECD, 2023).
6. DIF analysis ensures that mathematics tests fairly evaluate students from both urban and rural schools, different socio-economic backgrounds, and diverse language groups.

Differential Item Functioning in Rasch Modelling

Rasch models assume item invariance. When items fail to function equivalently across groups, DIF is present (Bond & Fox, 2015). Rasch-based DIF detection uses item difficulty contrasts, Wright maps, and item characteristic curves to visualize disparities (Linacre, 2020). Researches emphasize that graphical and statistical methods together provide more actionable insights for test developers (Wilson et al., 2017; Alderson et al., 2019). Odili and Okeke (2019) reported DIF effects in Nigerian secondary school mathematics tests. Chan et al. (2020) identified gender-related DIF in Malaysian secondary mathematics tests, while Retnawati et al. (2021) found DIF linked to language complexity in Indonesian mathematics assessments.

Differential Item Functioning (DIF) reinforced the necessity of examining items beyond aggregate statistics. The DIF analysis exposes whether items operate differently for diverse student groups, thereby safeguarding fairness in assessment outcomes (Zieky, 2021; Stoet & Geary, 2018; Odili & Okeke, 2019). Integrating Rasch modelling with DIF detection would

enable educators and researchers to identify and rectify biases arising from linguistic, cultural, gendered or instructional differences, ensuring that assessment scores reflect true mathematical ability rather than extraneous factors. This underscores the fact that rigorous test evaluation in mathematics requires a multi-faceted approach.

Statement of the Problem

Mathematics is one of the core and compulsory subjects for admission into tertiary institutions, especially in science, technology, engineering and mathematics (STEM) courses, yet many students struggle to do well in the subject. As observed by Adedoyin and Soyibo (2021), recent WAEC examinations show that a significant proportion of senior secondary school students in Nigeria, including those in Rivers State, fail to meet the minimum competency threshold, with persistently low pass rates that undermine their academic and professional prospects. Could this be attributed to faulty mathematics tests due to item misfit and poor differential item functioning (DIF)? Misfitting items can distort estimates of student performance, while items exhibiting DIF may unfairly favour certain groups, including male versus female students or students from urban versus rural schools (Zieky, 2021). Although the Rasch model has been widely acknowledged as a robust tool for addressing these limitations, its application in determining the item fit and DIF of senior secondary school mathematics tests in Rivers State remains limited, and this raises concern about the fairness, fitness and functioning of the kind of mathematics test used for critical decisions regarding student performance in the state. It was on this note that this study used the Rasch model to analyze the item fit and differential functioning of the senior secondary school mathematics test in Rivers State, Nigeria.

Purpose of the Study

The purpose of this study was to determine the fitness and differential functioning of senior secondary school mathematics test items in Rivers State. Specifically, the study sought to:

1. determine the fit of individual mathematics test items in senior secondary schools in Rivers State to the expectations of the Rasch model; and
2. examine the differential item functioning (DIF) of the mathematics test based on gender among Senior Secondary School students in Rivers State.

Research Questions

The study was guided by the following research questions:

1. To what extent do the senior secondary school mathematics test items in Rivers State demonstrate acceptable fit to the Rasch measurement model?
2. Do the senior secondary school mathematics test items in Rivers State demonstrate Differential Item

Functioning (DIF) by gender as measured by the Rasch model?

Hypotheses

The following null hypotheses were tested at a 0.05 level of significance:

1. The senior secondary school mathematics test items in Rivers State do not demonstrate acceptable fit to the Rasch model.
2. The senior secondary school mathematics test items in Rivers State do not demonstrate significant Differential Item Functioning (DIF) by gender as measured by the Rasch model.

Methods And Materials

The study adopted an instrumentation research design. An instrumentation research design is a special and important type of research design that is used for test development based on test theories to ensure a high level of satisfactory validity and reliability as well as an appropriate norm, criterion or domain for measuring and evaluating psychological attributes or human abilities (Kpolovie, 2010). This research design is suitable for the study, which aimed at analysing item fitness and differential functioning of senior secondary school mathematics test items in Rivers State using the Rasch model. The population of the study was 199,602 (97,198 male and 102,404 female students) in all the public senior secondary schools in Rivers State for the 2025/2026 academic session (Source: Rivers State Senior Secondary Schools Board, Port Harcourt, 2026). The sample size for the study was 480 students in public senior secondary schools in Rivers State. Taro Yamane's formula was used to obtain 399 samples, but the researcher decided to add 81 more students to make the sample 480 for better representation of the population and to ensure adequate male and female participation for item calibration, item fit analysis and Differential Item Functioning analysis by gender. Multi-stage sampling procedure was adopted for the sample selection. At the first stage, the population was stratified into 3 Senatorial Districts namely Rivers East, Rivers South East and Rivers West. At the second stage, 4 Local Government Areas (LGAs) were selected at random from each of the 3 Senatorial Districts given 12 LGAs. At the third stage, 2 public senior secondary schools were randomly selected from each of the 12 chosen LGAs given a total of 24 schools. At the fourth stage, 40 students (20 males and 20 females) were selected at random from each of the 24 schools, which gave a total sample of 480 students.

The instrument for data collection was a researcher-developed instrument titled: "Mathematics Performance Test for Senior Secondary School Two" (MPTSSS 2). The MPTSSS 2 is a multiple-choice test that contains 50 items with four response options lettered A–D, in which one is the correct answer, while the other three are plausible distractors. The researcher considered the content and objectives of the senior secondary school (SSS) II students' mathematics

syllabus in Rivers State and built a table of specification that guided the item generation; however, only four cognitive levels of Bloom's taxonomy were considered, which are knowledge, comprehension, application and evaluation, and four specific content areas such as Numbers and Numeration, Algebraic processes, Geometry, Trigonometry, Mensuration, and Statistics and Probability that are commonly assessed in school and external examinations. The MPTSSS 2 was scored dichotomously. The instrument was given to a Mathematics teacher in a public senior secondary school in Rivers State, and an expert in Measurement and Evaluation at Rivers State University, Port Harcourt, determined the face and content validity of the instrument by examining its clarity of language, suitability of instructions and appropriateness for SSS II students and made some modifications that guided revisions and the final copy. A table of specification was further developed to guide the generation of the items based on the Nigeria Senior Secondary School Mathematics curriculum, and to ensure adequate topic coverage, balanced cognitive levels, item relevance and proper representation of learning outcomes. Kuder–Richardson Formula 20 (KR-20) was used to calculate the internal consistency reliability coefficient using 30 SSS 2 students in one school in Rivers State, and a coefficient of 0.76 was obtained and considered high and reliable for the study.

Data collected were analysed using Rasch measurement procedures with the aid of the eRm package in R software. Gender was used as the focal grouping variable; hence, separate Rasch models were estimated for male and female students. Differential Item Functioning (DIF) analysis was also conducted to determine the item that functioned differently for male and female students. The difference between the male and female item-difficulty estimates was calculated using the formula:

$$DIF_i = b_{Mi} - b_{Fi}$$

Where:

DIF_i = DIF contrast for item i

b_{Mi} = Rasch item-difficulty estimate for item i among male students

b_{Fi} = Rasch item-difficulty estimate for item i among female students.

The standard error of the DIF contrast was calculated from the gender specific item standard errors using:

$$SE(DIF_i) = \sqrt{SE_{2Mi}^2 + SE_{2Fi}^2}$$

Where;

SE_{Mi} and SE_{Fi} represent respectively the standard errors of the male and female item-difficulty estimates.

The statistical significance of each DIF contrast was evaluated using the standardised z-statistic expressed as:

$$Z_i = \frac{DIF_i}{SE(DIF_i)}$$

Note that: $z > 1.96$ or $p < 0.05$ (DIF contrasts are statistically significant); $z \leq 1.96$ or $p \geq 0.05$ (DIF contrasts are not statistically significant). Rasch Criterion/Benchmark for decision: Item acceptable if infit MSQ and Outfit MSQ = 0.50 – 1.50. In terms of ethical considerations, express permission was obtained from the school principals before data collection from the students; voluntary participation was

sought from the students, while they were assured of the anonymity and confidentiality of their responses.

Results

Research Question 1: To what extent do the senior secondary school mathematics test items in Rivers State demonstrate acceptable fit to the Rasch measurement model?

Table 1a: Rasch Item Fit Statistics for the 50 Mathematics Test Items

Item	Infit MSQ	Outfit MSQ	Decision
Item 1	0.921	0.880	Acceptable
Item 2	0.882	0.842	Acceptable
Item 3	0.966	0.956	Acceptable
Item 4	0.921	0.880	Acceptable
Item 5	0.882	0.842	Acceptable
Item 6	0.949	0.907	Acceptable
Item 7	1.001	0.922	Acceptable
Item 8	0.966	0.956	Acceptable
Item 9	0.921	0.880	Acceptable
Item 10	1.229	1.392	Acceptable; but relatively high misfit
Item 11	0.921	0.880	Acceptable
Item 12	1.071	1.063	Acceptable
Item 13	0.994	0.986	Acceptable
Item 14	0.949	0.907	Acceptable
Item 15	0.969	0.987	Acceptable
Item 16	0.984	0.967	Acceptable
Item 17	0.923	0.899	Acceptable
Item 18	0.951	1.010	Acceptable
Item 19	1.071	1.063	Acceptable
Item 20	0.994	0.986	Acceptable
Item 21	0.933	0.895	Acceptable
Item 22	1.142	1.150	Acceptable
Item 23	1.001	0.922	Acceptable
Item 24	0.966	0.956	Acceptable
Item 25	1.145	1.424	Acceptable; but relatively high misfit
Item 26	0.923	0.899	Acceptable
Item 27	1.128	1.145	Acceptable
Item 28	1.107	1.104	Acceptable
Item 29	0.994	0.986	Acceptable
Item 30	0.977	0.967	Acceptable
Item 31	1.075	1.084	Acceptable
Item 32	0.972	0.944	Acceptable
Item 33	0.909	0.910	Acceptable
Item 34	0.933	0.895	Acceptable
Item 35	1.142	1.150	Acceptable
Item 36	0.971	0.931	Acceptable
Item 37	1.146	1.141	Acceptable
Item 38	0.949	0.907	Acceptable
Item 39	0.969	0.987	Acceptable
Item 40	1.074	1.326	Acceptable
Item 41	0.909	0.910	Acceptable
Item 42	1.132	1.127	Acceptable
Item 43	1.095	1.082	Acceptable
Item 44	0.949	0.907	Acceptable
Item 45	0.946	0.912	Acceptable

Item 46	0.984	0.979	Acceptable
Item 47	0.925	0.893	Acceptable
Item 48	1.057	1.093	Acceptable
Item 49	1.074	1.326	Acceptable
Item 50	0.909	0.910	Acceptable

Note: *Infit MSQ* = Mean Square; *Outfit MSQ* = Outfit Mean Square

Criterion: Values between 0.50 and 1.50 = acceptable item fit; above 1.50 = excessive unpredictability/noise; below 0.50 = very low (redundancy or excessive predictability).

Results in Table 1a showed that the entire 50-item senior secondary school mathematics test in Rivers State has Infit and Outfit MSQ values within the adopted acceptable range of 0.50–1.50. The Infit MSQ values range from 0.882 to 1.229, while the Outfit MSQ values range from 0.842 to 1.424. The results also showed that item 10 has the highest

Infit MSQ (1.229), while item 25 has the highest Outfit MSQ (1.424), but the values of both items 10 and 25 remain below the upper criterion of 1.50 and thus were not classified as misfitting or overfit items since they are within the adopted acceptable range of 0.50–1.50. It was concluded that all 50 items of the senior secondary school mathematics test demonstrated acceptable fit to the Rasch measurement model. However, the summary of the Rasch item-fit results is presented in Table 1b below.

Table 1b: Summary of Rasch Item-Fit Results

Indicator	Result	Interpretation
Number of items analysed	50	Complete test
Acceptable Infit MSQ	50 (100%)	All items fit
Acceptable Outfit MSQ	50 (100%)	All items fit
Highest Infit MSQ	1.229 (Item 10)	Acceptable
Highest Outfit MSQ	1.424 (Item 25)	Acceptable
Lowest Infit MSQ	0.882 (Items 2 & 5)	Acceptable
Lowest Outfit MSQ	0.842 (Items 2 & 5)	Acceptable
Items exceeding 1.50	0	None
Items below 0.50	0	None
Overall conclusion	50/50 fit	All items demonstrate acceptable Rasch fit

Table 1b showed that the entire 50 items have Infit and Outfit MSQ statistics that fall within the adopted criterion (acceptable) range (0.50–1.50), which clearly indicates that the senior secondary school mathematics test items conform adequately to the Rasch measurement model. In other words, there is no statistical basis from the MSQ fit statistics that can warrant deleting any of the 50 items solely because of item

misfit; hence, all the items demonstrated acceptable fit for the Rasch model.

Research Question 2: Do the senior secondary school mathematics test items demonstrate Differential Item Functioning (DIF) by gender as measured by the Rasch model?

Table 2: Rasch Model Analysis on DIF of Senior Secondary School Mathematics Test by Gender

Rank	Item	Male Difficulty	Female Difficulty	DIF Contrast	Interpretation
1	Item 18	0.525	0.706	-0.181	Largest DIF; favours males
2	Item 45	0.118	-0.047	0.165	Largest DIF; favours females
3	Item 48	0.098	0.254	-0.156	Substantial relative DIF; favours males
4	Item 15	0.326	0.192	0.134	Relative DIF; favours females
5	Item 39	0.326	0.192	0.134	Relative DIF; favours females
6	Item 36	-0.327	-0.200	-0.126	Relative DIF; favours males
7	Item 35	0.159	0.275	-0.116	Relative DIF; favours males
8	Item 22	0.159	0.275	-0.116	Relative DIF; favours males
9	Item 12	-0.001	-0.105	0.104	Relative DIF; favours females
10	Item 19	-0.001	-0.105	0.104	Relative DIF; favours females
11	Item 29	-0.156	-0.257	0.101	Relative DIF; favours females

12	Item 20	-0.156	-0.257	0.101	Relative DIF; favours females
13	Item 13	-0.156	-0.257	0.101	Relative DIF; favours females
14	Item 27	-0.345	-0.444	0.098	Relative DIF; favours females
15	Item 49	0.019	0.111	-0.092	Small-to-moderate difference
16	Item 40	0.019	0.111	-0.092	Small-to-moderate difference
17	Item 47	0.058	-0.028	0.086	Small difference
18	Item 17	-0.118	-0.047	-0.071	Small difference
19	Item 26	-0.118	-0.047	-0.071	Small difference
20	Item 25	0.118	0.051	0.067	Small difference
21	Item 31	-0.118	-0.066	-0.051	Minimal difference
22	Item 46	-0.156	-0.105	-0.051	Minimal difference
23	Item 32	-0.194	-0.143	-0.051	Minimal difference
24	Item 42	-0.098	-0.143	0.045	Minimal difference
25–50	Remaining items	—	—	≤ 0.036	Negligible/small DIF

Table 2 shows the Rasch analysis of DIF of the senior secondary school mathematics test by gender. The results revealed that the gender-related differences in item difficulty are generally small, with a few items that exhibited larger male–female contrasts. Item 18 showed the largest negative contrast (–0.181 logits), indicating greater estimated difficulty for females, while item 48 (–0.156 logits), items 15 and 39 (+0.134 logits each), item 36 (–0.126 logits), and items 22 and 35 (–0.116 logits each) also showed comparatively larger contrasts. However, the few items showing comparatively large contrasts are not enough to classify the entire test as demonstrating Differential Item

Functioning (DIF) by gender because the largest contrast is only about 0.18 logits, which is not strong or significant evidence of problematic DIF. Therefore, it was deduced that the senior secondary school mathematics test items do not demonstrate Differential Item Functioning (DIF) by gender as measured by the Rasch model.

Hypotheses Testing

Hypothesis 1: The senior secondary school mathematics test items in Rivers State do not demonstrate acceptable fit to the Rasch model.

Table 3: Diagnostic Analysis of Senior Secondary School Mathematics Test Items Demonstrating Acceptable Fit to the Rasch Model

Indicator	Result	Criterion	Decision
Number of items analysed	50	—	Complete test
Items with acceptable Infit MSQ	50 (100%)	0.50 –1.50	Acceptable
Items with acceptable Outfit MSQ	50 (100%)	0.50 –1.50	Acceptable
Highest Infit MSQ	1.229 (Item 10)	≤ 1.50	Acceptable
Highest Outfit MSQ	1.424 (Item 25)	≤ 1.50	Acceptable
Lowest Infit MSQ	0.882 (Items 2 & 5)	≥ 0.50	Acceptable
Lowest Outfit MSQ	0.842 (Items 2 & 5)	≥ 0.50	Acceptable
Items > 1.50	0 (0%)	0	None
Items < 0.50	0 (0%)	0	None
Overall item fit	50/50 (100%)	0.50 – 1.50	Acceptable

Note: MSQ = Mean Square; Infit MSQ = Information-Weighted Mean Square; Outfit MSQ = Outlier-Sensitive Mean Square.

Ho3 Rejected

The results in Table 3 show that all 50(100%) mathematics items had acceptable Infit and Outfit MSQ values, with their values falling within the criterion/benchmark (0.50-1.50), which is in line with the Rasch measurement model. Since the Infit and Outfit MSQ values fall within the criterion/benchmark of the Rasch measurement model, the null hypothesis three was therefore rejected; hence, it was concluded that the senior secondary school mathematics test

items in Rivers State demonstrated acceptable fit to the Rasch measurement model. This is because there was no statistical basis for deleting any of the mathematics test items.

Hypothesis 2: The senior secondary school mathematics test items do not demonstrate significant Differential Item Functioning (DIF) by gender as measured by the Rasch model.

Table 4: Rasch Differential Item Functioning Statistics for Male and Female Students

Item	Male Difficulty	Female Difficulty	DIF Contrast	DIF SE	Z	p-value	Decision
Item 1	0.1183	0.0906	0.0277	0.1983	0.140	.8889	No Significant DIF
Item 2	-0.2323	-0.2003	-0.0320	0.1928	-0.166	.8683	No Significant DIF
Item 3	0.1998	0.2122	-0.0124	0.2010	-0.062	.9508	No Significant DIF
Item 4	0.1183	0.0905	0.0277	0.1983	0.140	.8888	No Significant DIF
Item 5	-0.2323	-0.2003	-0.0320	0.1928	-0.166	.8683	No Significant DIF
Item 6	-0.4392	-0.4439	0.0047	0.1914	0.024	.9806	No Significant DIF
Item 7	0.5253	0.5618	-0.0365	0.2128	-0.171	.8640	No Significant DIF
Item 8	0.1998	0.2122	-0.0124	0.2010	-0.062	.9508	No Significant DIF
Item 9	0.1183	0.0905	0.0277	0.1983	0.140	.8888	No Significant DIF
Item 10	-0.5328	-0.4997	-0.0331	0.1914	-0.173	.8627	No Significant DIF
Item 11	0.1183	0.0905	0.0277	0.1983	0.140	.8888	No Significant DIF
Item 12	-0.0011	-0.1050	0.1039	0.1952	0.532	.5946	No Significant DIF
Item 13	-0.1561	-0.2569	0.1008	0.1930	0.523	.6013	No Significant DIF
Item 14	-0.4392	-0.4439	0.0047	0.1914	0.024	.9806	No Significant DIF
Item 15	0.3258	0.1917	0.1342	0.2026	0.662	.5078	No Significant DIF
Item 16	0.1588	0.1509	0.0079	0.1996	0.040	.9683	No Significant DIF
Item 17	-0.1177	-0.0471	-0.0706	0.1947	-0.363	.7168	No Significant DIF
Item 18	0.5253	0.7060	-0.1807	0.2162	-0.836	.4034	No Significant DIF
Item 19	-0.0011	-0.1050	0.1039	0.1952	0.532	.5946	No Significant DIF
Item 20	-0.1561	-0.2569	0.1008	0.1930	0.523	.6013	No Significant DIF
Item 21	-0.2323	-0.2381	0.0058	0.1926	0.030	.9759	No Significant DIF
Item 22	0.1588	0.2747	-0.1159	0.2014	-0.575	.5650	No Significant DIF
Item 23	0.5253	0.5618	-0.0365	0.2128	-0.171	.8640	No Significant DIF
Item 24	0.1998	0.2122	-0.0124	0.2010	-0.062	.9508	No Significant DIF
Item 25	0.1183	0.0508	0.0675	0.1979	0.341	.7331	No Significant DIF
Item 26	-0.1177	-0.0471	-0.0706	0.1947	-0.363	.7168	No Significant DIF
Item 27	-0.3455	-0.4439	0.0984	0.1916	0.514	.6075	No Significant DIF
Item 28	-0.0011	0.0114	-0.0125	0.1962	-0.064	.9492	No Significant DIF
Item 29	-0.1561	-0.2569	0.1008	0.1930	0.523	.6013	No Significant DIF
Item 30	0.1385	0.1307	0.0078	0.1991	0.039	.9686	No Significant DIF
Item 31	-0.1177	-0.0665	-0.0512	0.1945	-0.263	.7922	No Significant DIF
Item 32	-0.1943	-0.1433	-0.0510	0.1934	-0.264	.7920	No Significant DIF
Item 33	0.3258	0.3384	-0.0126	0.2048	-0.061	.9511	No Significant DIF
Item 34	-0.2323	-0.2381	0.0058	0.1926	0.030	.9759	No Significant DIF
Item 35	0.1588	0.2747	-0.1159	0.2014	-0.575	.5650	No Significant DIF
Item 36	-0.3267	-0.2003	-0.1264	0.1925	-0.657	.5114	No Significant DIF
Item 37	-0.1369	-0.1433	0.0063	0.1938	0.033	.9739	No Significant DIF
Item 38	-0.4392	-0.4439	0.0047	0.1914	0.024	.9806	No Significant DIF
Item 39	0.3258	0.1917	0.1342	0.2026	0.662	.5078	No Significant DIF
Item 40	0.0186	0.1105	-0.0920	0.1975	-0.466	.6414	No Significant DIF
Item 41	0.3258	0.3384	-0.0126	0.2048	-0.061	.9511	No Significant DIF
Item 42	-0.0985	-0.1433	0.0448	0.1940	0.231	.8173	No Significant DIF
Item 43	-0.0207	-0.0471	0.0264	0.1955	0.135	.8926	No Significant DIF
Item 44	-0.4392	-0.4439	0.0047	0.1914	0.024	.9806	No Significant DIF
Item 45	0.1183	-0.0471	0.1654	0.1969	0.840	.4010	No Significant DIF
Item 46	-0.1561	-0.1050	-0.0511	0.1939	-0.264	.7921	No Significant DIF
Item 47	0.0582	-0.0277	0.0859	0.1964	0.437	.6620	No Significant DIF
Item 48	0.0981	0.2537	-0.1556	0.2003	-0.777	.4373	No Significant DIF
Item 49	0.0186	0.1105	-0.0920	0.1975	-0.466	.6414	No Significant DIF
Item 50	0.3258	0.3384	-0.0126	0.2048	-0.061	.9511	No Significant DIF

Note: DIF = Differential Item Functioning; SE = Standard Error; Positive values indicate greater estimated difficulty for male students; Negative values indicate greater estimated difficulty for female students; DIF was evaluated at $\alpha = 0.05$, two-tailed test.

The results of Rasch DIF analysis, as shown in Table 4, revealed that at the 0.05 level of significance, the p-values for all 50 mathematics items ranged from 0.4010 to 0.9806, which are higher than the 0.05 level of significance; hence, none of the items demonstrated statistically significant Differential Item Functioning (DIF) by gender. Therefore, it was concluded that the senior secondary school mathematics test does not demonstrate significant Differential Item Functioning by gender. This result implies that the performance of the students on the test was not based on gender.

Discussion

Item Fit (or Fitness) to the Rasch Measurement Model

The analysis of data for Research Question 1 and its corresponding Null Hypothesis 1 revealed that all the items of the senior secondary school mathematics test demonstrated acceptable fit to the Rasch measurement model. This implies that all 50 mathematics items used for the study generally showed Infit and Outfit values that are generally close to the expected value of 1.00 and are consistent with the Rasch model. The result of this study is critical in the process of developing mathematics assessment instruments because of the notion that a test may demonstrate satisfactory reliability even though some of the individual items may fail to function in accordance with the underlying Rasch measurement model. This could be because the items were properly scrutinised and reviewed by experts during the drafting stage. In addition, Chan et al. (2020) explained that misfitting mathematics items can reduce measurement precision among secondary-school students, and thus emphasised the importance of identifying items whose response patterns depart substantially from model expectations. This is to ensure that the items possess acceptable fit to contribute positively to the Rasch measurement model without introducing substantial unexplained response variation. Ojerinde (2020) noted that misfitting items in Nigerian mathematics examinations may arise from ambiguous phrasing or culturally unfamiliar contexts, but item-fit analysis can strengthen construct validity, improve reliability, enhance fairness and guide test development. This is important because mathematics tests administered to senior secondary school students in Nigeria are used for important educational decisions, including promotion, selection and certification; hence, a misfit item could potentially affect the results and their interpretation. In slight disagreement, Retnawati et al. (2021) argued that approximately 10–15% of mathematics items exhibited misfit associated with ambiguity and cognitive overload. This calls for careful test construction and psychometric evaluation as to reduce some of the

problems associated with traditionally developed mathematics tests in Nigeria.

Differential Item Functioning of the Senior Secondary School Mathematics Test by Gender

The results for Research Question 2 and its Null Hypothesis 2 indicated that the senior secondary school mathematics test items do not demonstrate Differential Item Functioning (DIF) by gender as measured by the Rasch model. This result is because the observed gender-related differences in item difficulty are generally small, with very few items showing comparatively large male–female contrast, which are not enough to classify the entire test as demonstrating Differential Item Functioning (DIF) based on gender; the largest contrast is only about 0.18 logits, and it is not significant evidence of problematic DIF. In support, Kaya and Demir (2021), who in their study focused on detecting gender-based DIF in high school mathematics tests in Turkey using the Rasch model, reported that seven items exhibited gender-based DIF without any systematic favouring of either the male or female students. Njeri and Mwangi (2019) examined gender DIF in secondary school mathematics assessments using Rasch Modelling and found that only six items demonstrated significant DIF, which was not enough to conclude that there existed significant DIF. In contrast, Kim and Lee (2023) reported gender-related DIF in South Korean secondary mathematics assessments. However, it should be noted that gender-related DIF is not a universal feature of any nation's test; its occurrence in a test depends on a particular characteristic of the items and the students who responded to them. It is also worthy to reiterate that the absence of statistically significant gender DIF in a mathematics test does not mean that gender has no relationship whatsoever with the mathematics achievement, it is because DIF and group performance differences are conceptually different.

Implications of the Findings

With the findings of the study, test developers and examination bodies would begin to consider conducting empirical analysis of item functioning seriously using the Rasch model in order to retain those that showed acceptable fit and contribute meaningfully to the measurement continuum instead of retaining items solely because they appear content-relevant. Also, the findings would indirectly give students a fairer opportunity to be assessed with test items that are neither too easy nor too difficult, with no substantial or systematic gender based DIF that would undermine the fair assessment and measurement.

Recommendations

Based on the findings of the study, the following recommendations were made:

1. Mathematics teachers should collaborate with measurement experts for use of the Rasch model for identification of potential item misfits for review or

replacement to ensure that all items demonstrate acceptable fit for retention in the final test.

2. Schools' Continuous Assessment/ Examination Committees should always review mathematics test items for Differential Item Functioning (DIF) by gender to be satisfied that the test does not introduce gender-related bias before administering it to the students.

Contributions to Knowledge

In terms of mathematics test development, the study has shown how the findings from the Rasch measurement model can be transformed into practical or evidence-based information for mathematics test development, validation and improvement. The findings about item misfit and differential functioning serve as an eye-opener to test developers on the need to use the Rasch model for review and validation of test items in order to ascertain which items qualify for retention or removal in a test bank. In addition, the study has contributed to the growing body of empirical evidence that supports the use of modern measurement approaches such as the Rasch model instead of relying on the traditional approach like classical test methods in the Nigerian educational system.

References

1. Adedoyin, O. O., & Soyibo, K. (2021). Mathematics achievement and assessment practices in secondary schools. *African Journal of Educational Research*, 25(2), 112–129.
2. AERA, APA, & NCME. (2018). *Standards for educational and psychological testing*. American Educational Research Association.
3. Alderson, J. C., Brunfaut, T., & Harding, L. (2019). Language complexity and DIF in educational assessments. *Language Testing*, 36(3), 345–364.
4. Aminu, I. A., & Bello, M. S. (2019). Rasch analysis of senior secondary school mathematics items in Kaduna State, Nigeria. *Nigerian Journal of Educational Measurement*, 11(1), 55–72.
5. Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (2nd ed.). Springer. <https://doi.org/10.1007/978-94-007-6852-5>
6. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315742972>
7. Chan, S. W., Ismail, Z., & Sumintono, B. (2020). Rasch model applications in mathematics education assessment. *International Journal of Instruction*, 13(1), 247–266.
8. De Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.
9. DeMars, C. E. (2020). *Item response theory and psychometrics in education*. Routledge.
10. DeMars, C. E. (2020). *Measurement and psychometrics in educational testing*. Routledge.
11. Federal Republic of Nigeria (RGN, 2014). *National policy on education*. NERDC.
12. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2016). *Fundamentals of item response theory* (2nd ed.). Sage.
13. Hidalgo, M. D., Gómez-Benito, J., & Padilla, J. L. (2020). Differential item functioning detection in educational testing. *Frontiers in Psychology*, 11, 1–12.
14. Ismail, Z., & Sumintono, B. (2021). Rasch analysis and item bias in mathematics assessment. *Educational Measurement Journal*, 18(1), 15–34.
15. Kaya, Y., & Demir, H. (2021). Detecting gender-based DIF in high school mathematics tests using the Rasch model. *Eurasian Journal of Educational Measurement*, 11(3), 112–130.
16. Kim, H., & Lee, J. (2023). Gender differential item functioning in South Korean mathematics assessments using Rasch analysis. *Asia-Pacific Education Researcher*, 32(1), 15–28. <https://doi.org/10.1007/s40299-022-00674-3>
17. Kpolovie, P. J. (2010). *Advanced Research Methods*. Springfield Publishers Ltd.
18. Linacre, J. M. (2020). *Winsteps Rasch measurement computer program manual*. Winsteps.com.
19. Mwangi, J., Kibet, S., & Otieno, P. (2020). Item fit analysis of secondary school mathematics tests using Rasch modelling. *African Journal of Educational Evaluation*, 12(2), 74–92.
20. Njeri, P., & Mwangi, J. (2019). Examining gender DIF in secondary school mathematics assessments using Rasch modelling. *Journal of Educational Evaluation in Africa*, 14(2), 57–76.
21. Ojerinde, D. (2020). Educational assessment and examination integrity in Nigeria. *Nigerian Journal of Educational Measurement*, 21(1), 15–29.
22. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. (Original work published 1960)
23. Retnawati, H., Hadi, S., & Zubaidah, S. (2021). Improving mathematics assessment using Rasch measurement models. *Journal of Mathematics Education*, 14(2), 25–38.
24. Smith, T., & Johnson, R. (2021). IRT and Rasch modelling in high school mathematics assessment: Improving predictive validity. *Journal of Educational Measurement*, 58(1), 33–50. <https://doi.org/10.1111/jedm.12223>
25. Zieky, M. (2021). Detecting and evaluating differential item functioning in educational tests. *Applied Measurement in Education*, 34(2), 101–115.