



Evaluation of the Rasch Model in Determining the Reliability and Difficulty Indices of Senior Secondary School Mathematics Tests in Rivers State

EJIMAJI, F. U.¹; OBILOR, E. I.²; & MIWARI, G. U.³

Department of Educational Foundations, Faculty of Education, Rivers State University, Port Harcourt

Article history:

Received: 09/07/2026

Accepted: 18/08/2026

Published: 05/10/2026

Keywords: Difficulty level, item reliability, person reliability, Rasch model, test

*Corresponding Author:

EJIMAJI, F. U.

Abstract

The study evaluated the Rasch Model in determining the reliability and difficulty indices of senior secondary school mathematics tests in Rivers State. With the instrumentation design, the study specifically examined person reliability, item reliability and the difficulty level of the items. Three research questions and associated null hypotheses guided the study. From a population of 199,602 students in the public senior secondary schools in Rivers State in the 2025/2026 academic session, a sample of 480 Senior Secondary Two (SS2) students was selected for the study through a multi-stage sampling procedure. "Mathematics Performance Test (MPT) for SS2 Students" was the instrument used for data collection. The MPT was validated by Mathematics and Evaluation experts. The KR₂₀ was used to establish the internal consistency reliability of the instrument, and a reliability coefficient of 0.76 was obtained. Data was analysed using the Rasch model with the aid of the eRm package in R software. Results of the study showed that the senior secondary school mathematics tests in Rivers State possessed high person reliability, good item reliability, and provided a reasonably varied range of item difficulties. It was recommended that mathematics test developers should utilise the Rasch model for calibration of mathematics test items to ensure that all the items attain high reliability coefficients to dependably measure the students' mathematics abilities, and that mathematics teachers should adopt mathematics tests that contain a varied range of easy, moderate and difficult items to create a level playing field for both low and high scorers.

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., & MIWARI, G. U. (2026). Evaluation of the Rasch model in determining the reliability and difficulty indices of senior secondary school mathematics tests in Rivers State. *EIRA Journal of Arts, Law and Educational Sciences*, 2(5). 28-37.

Introduction

Mathematics is the building stone, or foundation, on which scientific and technological innovations are built in a society. In school, especially at the primary and secondary levels, all pupils/students study mathematics as a core subject because it develops learners' mental ability and reasoning. To prove its importance, the subject is a prerequisite for admission to tertiary institutions for all science-based courses such as engineering, medicine, computer science, architecture, and environmental sciences, as well as almost all social science and business-based courses. The study of mathematics is so important because it builds analytical reasoning, problem-solving abilities, and logical reasoning skills that are needed to navigate the twenty-first century tasks and career demands. This implies that the quality of mathematics teaching, as well as the performance of students, has great consequences for the students' academic progression, career prospects and overall national development. In Nigeria, it has long been a policy that the curriculum framework emphasises the need for

students to acquire quantitative reasoning, logical thinking, and analytical skills necessary for national growth and development (Federal Republic of Nigeria, 2014). To achieve this, educational institutions and stakeholders must pay close attention to the psychometric qualities of mathematics tests and other instruments used to measure students' learning outcomes. Psychometric qualities refer to the characteristics of a test that determine its quality, usefulness, accuracy, and appropriateness for measurement purposes. This is because the quality of any educational assessment largely depends on the psychometric properties of the test instrument used (Retnawati et al., 2021). Any educational instrument that has poor psychometric quality is highly prone to produce misleading indices and thus negatively affect educational decisions made based on the results. It is on this premise that this study sought to determine the reliability indices and difficulty levels of senior secondary school mathematics tests in Rivers State using the Rasch model.

The Rasch Model

Rasch model is now one of the most influential developments in modern educational measurement and psychometrics, which has transformed the way researchers conceptualise testing, measurement precision, item functioning and ability estimation of instruments. Bond and Fox (2015) explained that in educational assessment, especially within mathematics education, the Rasch model is increasingly recognised as a scientifically rigorous framework capable of addressing many of the limitations associated with traditional measurement approaches like Classical Test Theory (CTT). The Rasch model is concerned with the construction of objective measurement systems capable of producing invariant estimates of both item difficulty and person ability. Rasch model is unique in that it has the capacity to transform ordinal observations into interval-level measures, thereby enabling more meaningful statistical interpretation and comparison, unlike other conventional models (Embretson & Reise, 2018). In fact, the Rasch model provides a robust framework for evaluating whether mathematics test items function appropriately across diverse student populations, including students from urban and rural schools, different socioeconomic backgrounds, and varying instructional contexts. The Rasch model aligns with contemporary educational philosophies that emphasise fairness, evidence-based assessment, and learner-centred evaluation. Rather than viewing assessment merely as a mechanism for ranking students, Rasch measurement supports the development of assessments that accurately reflect students' abilities and learning needs (Brookhart & Nitko, 2019). Through detailed item-level diagnostics, the model enables researchers to identify poorly functioning items, biased questions, and inconsistencies in test design.

Reliability

Reliability refers to the consistency, stability, and dependability of test scores. Reliability concerns the consistency and stability of assessment outcomes across repeated administrations or different forms of the test (Retnawati et al., 2021). Reliability is fundamental because unreliable assessments cannot produce meaningful interpretations. Reliability is one of the most fundamental concepts in educational measurement and psychometric evaluation because it determines the degree to which assessment outcomes are stable, dependable, and consistent across different testing situations. In educational assessment, particularly in mathematics testing, reliability is indispensable because decisions concerning students' promotion, placement, certification, and academic progression are often based on examination scores. Contemporary psychometric theory conceptualises reliability more broadly as the extent to which observed scores are free from random measurement errors (De Ayala, 2022). This implies that reliability is not merely a property of a test itself, but rather a characteristic of test scores generated within specific contexts and populations.

In mathematics education, reliability is particularly important because mathematical competence involves several interconnected dimensions, including procedural fluency, conceptual understanding, logical reasoning, problem-solving ability, and computational accuracy (Kilpatrick et al., 2021). In Nigeria, concerns regarding the reliability of senior secondary school mathematics examinations have intensified due to persistent fluctuations in students' performance in public examinations such as the West African Senior School Certificate Examination [by the West African Examinations Council] and the Senior School Certificate Examination [by the National Examinations Council] (Ojerinde, 2020), which could be attributed to ambiguous wording and inconsistencies in scoring procedures; these concerns underscore the importance of psychometric evaluation in ensuring that mathematics assessments produce dependable and interpretable scores.

Reliability (item and person) using the Rasch Model

Rasch model reliability provides more meaningful psychometric information than traditional reliability coefficients because it incorporates both item difficulty and person ability simultaneously (Retnawati et al., 2021). In Rasch model analysis, reliability is classified into person reliability and item reliability.

Person reliability: Person reliability indicates how consistently the test distinguishes among students with varying ability levels. Person reliability provides a quantitative lens through which one can evaluate whether a mathematics test meaningfully distinguishes between high-performing and low-performing students. Unlike classical reliability indices, such as Cronbach's alpha, person reliability is grounded in probabilistic measurement, emphasising the stability of ability estimates derived from the test (Linacre, 2022). A high person reliability coefficient suggests that a test has sufficient breadth and depth to capture variations in student ability; low person reliability, on the other hand, indicates that items are too homogeneous or misaligned with the ability spectrum of the target population. Wang and Wilson (2023) examined the secondary mathematics assessments in Singapore and observed a person reliability coefficient above 0.90, which substantially improved the accuracy of student classification; hence, they concluded that it was crucial in high-stakes contexts where promotion and certification depend on test outcomes. Similarly, Yusuf and Alabi (2022) reported person reliability of 0.84 in Nigerian mathematics assessments, highlighting that even moderate reliability can yield meaningful differentiation if coupled with careful test design. The concept extends beyond statistical consistency. It embodies fairness and interpretability: when tests accurately reflect differences in ability, stakeholders including educators, students, and policymakers can trust the outcomes and make informed decisions. Reliable mathematics assessments minimise measurement error and produce dependable estimates of students' ability. Modern psychometric

approaches such as Item Response Theory (IRT) and Rasch analysis provide more sophisticated reliability estimates than traditional methods because they simultaneously evaluate item difficulty and person ability (Bond & Fox, 2015).

Item reliability: Item reliability refers to the consistency and stability of item difficulty estimates across different samples of respondents. Item reliability complements person reliability by focusing on the consistency and stability of item difficulty across samples. In Rasch analysis, item reliability indicates the extent to which the hierarchy of item difficulty would remain stable if the same items were administered to another comparable sample of students. Item reliability is therefore a critical indicator of the quality and reproducibility of test items. Boone et al. (2014) explained that item reliability reflects the degree of confidence that can be placed in item calibrations. High item reliability suggests that item difficulty estimates are stable and can be generalised beyond the sample used in the analysis (Linacre, 2022). On the other hand, low item reliability indicates insufficient sample size, narrow ability distribution, or inconsistent item functioning. Item reliability values range between 0 and 1, and any value above 0.80 is regarded as high for educational tools, and a value of 0.90 and above is taken as excellent item calibration stability.

Item reliability is closely related to item separation index because high item separation index implies that the test contains a sufficiently wide range of easy, moderate, and difficult items, while high item reliability ensures that the hierarchy of items ranging from easy to challenging remains reproducible, a principle critical for fair assessment and meaningful comparison across student cohorts (Boone et al., 2014). Adeyemi and Lawal (2021) demonstrated the importance of item reliability in educational assessment when they evaluated Senior Secondary School Mathematics test items in Osun State, Nigeria, using Rasch analysis and reported item reliability coefficients above 0.92, and thus concluded that high item reliability contributed significantly to the validity and stability of mathematics achievement scores. Hassan and Alshamrani (2021) investigated secondary school mathematics assessments in Saudi Arabia and found that Rasch item reliability improved confidence in score interpretation and item calibration. Item reliability is particularly important in standardised examinations such as the West African Senior School Certificate Examination and the Senior School Certificate Examination because stable item calibrations contribute to fairness and comparability across examination years.

Reliability in Classical Test Theory

Under Classical Test Theory, reliability is primarily estimated using coefficients such as: test-retest reliability, parallel forms reliability, internal consistency reliability, and inter-rater reliability.

Test-Retest Reliability: Test-retest reliability determines the stability of scores over time when the same test is

administered to the same group on two different occasions. The scores are correlated.

Parallel Forms Reliability: Parallel form reliability involves administering two equivalent versions of a test to the same group of students. The correlation between scores indicates consistency across forms (De Ayala, 2022). This approach is particularly relevant in national examinations where multiple versions of test papers may be used to reduce examination malpractice.

Internal Consistency Reliability: Internal consistency examines the degree to which items within a test measure the same construct. It is among the most commonly used reliability indices in educational research. Internal consistency is measured by Cronbach's Alpha, which is widely used to estimate internal consistency reliability. Values closer to 1.00 indicate higher reliability, and split-half reliability also measures internal consistency. Test items are divided into two equivalent halves, and the scores are correlated. The Spearman-Brown prophecy formula is usually used to estimate the total reliability.

Inter-Rater Reliability: Inter-rater reliability concerns the consistency of scores among raters. This form is especially relevant in mathematics assessments involving essay questions or extended problem-solving. Poor inter-rater reliability can compromise fairness and lead to inconsistent evaluation outcomes.

Difficulty Levels of Mathematics Test Items on the Rasch Scale

Okonkwo and Eze (2018), in their study which aimed at determining the spread of item difficulties and their alignment with students' abilities, indicated that while most items were appropriately distributed, the test lacked extremely difficult items to challenge high-ability students. Wanjiku and Kamau (2019) examined Rasch-based analysis of secondary school mathematics test items, examining difficulty levels and student performance in Kenya, and reported that items ranged from -2.45 logits for the easiest item to 2.50 logits for the most difficult, with a clustering of items in the medium difficulty range.

Lim and Tan (2020) examined mathematics test item difficulty estimation using Rasch analysis in Singapore and revealed item difficulty estimates ranging from -2.30 to 2.40 logits, showing a good spread but indicating that items at the extremes of the ability scale were underrepresented. In the same vein, Kaya and Aydin (2021) in their study which investigated the difficulty levels of test items for accurate measurement of student abilities. The results indicated that item difficulty spanned from -2.20 to 2.45 logits, with an overrepresentation of medium-difficulty items. The study recommended revising the test to include more items at the extreme difficulty levels to enhance discrimination. Further, Dlamini and Nkosi (2022), in their investigation of difficulty levels and student performance in South Africa, found that most items were clustered around moderate difficulty, with

few items targeting the extremes of abilities. This calls for the inclusion of lower- and higher-difficulty items for better assessment coverage. The above studies complement the present research by providing evidence of the importance of item difficulty calibration in high school mathematics assessment.

Statement of the Problem

Mathematics at the senior secondary school (SSS) level is a core and compulsory subject for admission into tertiary institutions, particularly in science, technology, engineering, and mathematics (STEM) disciplines (FME, 2014), yet recent WAEC examination results show that a significant proportion of SSS students in Nigeria, including those in Rivers State, fail to meet the minimum competency threshold, with persistent low pass rates undermining their academic and professional prospects (Adedoyin & Soyibo, 2021). The quality of assessment instruments has been identified as one of the factors in poor performance of students in mathematics. This is because mathematics tests are intended to provide accurate, fair, and interpretable measures of students' abilities, yet many existing mathematics assessments are developed and evaluated using Classical Test Theory (CTT) with its inherent limitations. Psychometric evaluation of senior secondary school mathematics tests tends to overlook critical issues such as alignment between item difficulty and student ability, and the person reliability of the test. Although the Rasch model has been widely acknowledged as a robust tool for addressing these limitations, its application to senior secondary school mathematics tests in Rivers State remains scanty, which has created a huge gap in the literature for the conduct of this study. This study therefore focused on the evaluation of the Rasch Model in determining the reliability and difficulty indices of senior secondary school mathematics tests in Rivers State.

Purpose of the Study

The purpose of the study was to evaluate the Rasch Model in determining the reliability and difficulty indices of senior secondary school mathematics tests in Rivers State. Specifically, the study sought to:

1. determine the person reliability of senior secondary school mathematics tests in Rivers State using the Rasch model;
2. examine the item reliability of senior secondary school mathematics tests; and
3. estimate the difficulty levels of mathematics test items on a common Rasch scale.

Research Questions

The study was guided by the following research questions:

1. What is the level of person reliability of the senior secondary school mathematics tests in Rivers State?

2. To what extent does the senior secondary school mathematics test possess good item reliability as measured using the Rasch model?
3. What are the difficulty levels of the senior secondary school mathematics test items in Rivers State as measured on the Rasch scale?

Hypotheses

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

1. The senior secondary school mathematics test does not demonstrate a high level of person reliability as measured by the Rasch model.
2. The senior secondary school mathematics test does not demonstrate adequate item reliability as measured using the Rasch model.

Methods And Materials

The study adopted an instrumentation research design to evaluate the Rasch Model in determining the reliability and difficulty indices of senior secondary school mathematics tests in Rivers State. Kpolovie (2010) viewed instrumentation research design as a special and important type of research design that is primarily 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, which is the focus of the current study. The population of the study was 199,602 students, made up of 97,198 males and 102,404 females 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). Four hundred and eighty (480) Senior Secondary Two (SS2) students were sampled for the study. Taro Yamane's formula was used to arrive at a sample size of 399, but this was increased to 480 to ensure adequate and equal representation of male and female participants, and good item calibration. A multi-stage sampling procedure was adopted in which, at the first stage, the population was stratified into three Senatorial Districts of Rivers East, Rivers South-East and Rivers West. At the second stage, four Local Government Areas (LGAs) were selected at random in each of the three Senatorial Districts, giving 12 LGAs. At the third stage, two public senior secondary schools were randomly selected from each of the 12 chosen LGAs, and at the fourth stage, 40 students, made up of 20 males and 20 females, were selected at random from each of the 24 schools, which gave a sample size of 480 SS2 students. A researcher-developed instrument titled Mathematics Performance Test for Senior Secondary School Two" (MPTSSS 2) was used for data collection. The MPTSSS 2 comprises 50-multiple-choice test items with four options lettered A–D, in which one is the key, while the other three are the distractors. The test covered four content areas – Numbers and Numeration; Algebraic Processes; Geometry, Trigonometry, and Mensuration; and Statistics and Probability – which were dichotomously scored. The face

and content validities of the instrument were determined by three experts: a Mathematics teacher in a public senior secondary school in Rivers State, and two lecturers in the area of Measurement and Evaluation at Rivers State University, Port Harcourt, by examining its clarity of language, suitability of instructions and appropriateness for SS2 students and moderating a few of the items. The table of specifications was further used to guide the generation of the items to ensure good coverage and adequate representation of learning outcomes. The instrument was first administered to 30 SS2 students in a school that was not selected for the study, and their scores were calculated for the internal consistency

reliability coefficient using Kuder–Richardson Formula 20 (KR-20); a reliability coefficient of 0.76 was obtained, which was considered high and suitable for the study. The data collected was analysed using Rasch measurement procedures with the aid of the eRm package in R software. Rasch Criterion/Benchmark for decision: High/good Person Reliability = 0.70; high/good Item Reliability $\geq$ 0.80.

Results

Research Question 1: What is the level of person reliability of the senior secondary school mathematics tests in Rivers State?

Table 1: Rasch Person Reliability Statistics of Senior Secondary School Mathematics Tests in Rivers State

Statistic	Value	Interpretation
Number of respondents	480	Complete sample
Persons with estimated ability parameters	459	95.6% of respondents
Extreme persons	21	4.4% of respondents
Variance of person ability (SD^2)	0.7884	Meaningful spread in student abilities
Mean error variance (SE^2)	0.1143	Relatively small measurement error
Person reliability (R_p)	0.8550	High/good reliability
Person separation index	2.4286	Good separation of persons
Approximate number of ability strata	3.57	Approximately 3–4 distinguishable ability levels

Table 1 reveals a person reliability coefficient of 0.855, a person separation index of 2.429, with 459 (95.6%) respondents having estimated ability parameters, while 21 (4.4%) respondents were

identified as extreme persons; however, this does not invalidate the overall person reliability result because the extreme responses were handled by the Rasch estimation procedure through interpolation in the person parameter table. The obtained person reliability of 0.8550 was considered satisfactory/high for the purpose of differentiating

respondents on the measured construct. Hence, it was concluded that the level of person reliability of the senior secondary school mathematics test in Rivers State is high. This implies that the test was adequate and sufficiently consistent in distinguishing students according to their estimated levels of mathematical ability.

Research Question 2: What is the extent to which the senior secondary school mathematics test possesses good item reliability as measured using the Rasch model?

Table 2: Rasch Item Reliability Statistics of the Senior Secondary School Mathematics Test

Statistic	Value	Interpretation
Number of test items	50	Complete test
Variance of item difficulty estimates	0.0714	Adequate spread of item difficulties
Mean squared standard error of item estimates	0.00976	Low measurement error
Item reliability	0.8633	High/good reliability
Item separation index	2.5131	Good item separation
Approximate number of item strata	3.68	Approximately 4 distinguishable difficulty levels

Table 2 presents the Rasch analysis of item reliability of the senior secondary school mathematics test. The results showed that the variance of the item difficulty estimates = 0.0714, mean squared standard error of item = 0.00976, item reliability coefficient = 0.8633 and item separation index = 2.5131. The item separation index of 2.5131 implies that the test contains a sufficiently wide range of item difficulties to permit meaningful ordering of the items, while the item reliability coefficient of 0.863 indicates that the ordering of

the 50 mathematics test items according to their difficulty is highly reliable; hence, it was concluded that, to a high extent, the senior secondary school mathematics test possesses good item reliability.

Research Question 3: What are the difficulty levels of the senior secondary school mathematics test items in Rivers State as measured on the Rasch scale?

Table 3a: Rasch Item-Difficulty Estimates of the Senior Secondary School Mathematics Test

Rank	Item	Proportion- Correct	Rasch Difficulty (logits)	Interpretation
1	Item 18	0.746	-0.614	Very easy
2	Item 23	0.733	-0.544	Very easy
3	Item 7	0.733	-0.544	Very easy
4	Item 50	0.694	-0.332	Easy
5	Item 33	0.694	-0.332	Easy
6	Item 41	0.694	-0.332	Easy
7	Item 15	0.679	-0.258	Easy
8	Item 39	0.679	-0.258	Easy
9	Item 35	0.671	-0.217	Easy
10	Item 22	0.671	-0.217	Easy
11	Item 24	0.669	-0.206	Easy
12	Item 8	0.669	-0.206	Easy
13	Item 3	0.669	-0.206	Easy
14	Item 48	0.663	-0.175	Easy
15	Item 16	0.658	-0.155	Easy
16	Item 30	0.654	-0.135	Easy
17	Item 4	0.648	-0.105	Easy
18	Item 9	0.648	-0.105	Easy
19	Item 11	0.648	-0.105	Easy
20	Item 1	0.648	-0.105	Easy
21	Item 25	0.644	-0.084	Moderate
22	Item 49	0.640	-0.065	Moderate
23	Item 40	0.640	-0.065	Moderate
24	Item 45	0.633	-0.035	Moderate
25	Item 47	0.629	-0.015	Moderate
26	Item 28	0.627	-0.005	Moderate
27	Item 43	0.619	0.034	Moderate
28	Item 12	0.615	0.053	Moderate
29	Item 19	0.615	0.053	Moderate
30	Item 17	0.608	0.082	Moderate
31	Item 26	0.608	0.082	Moderate
32	Item 31	0.606	0.092	Moderate
33	Item 42	0.600	0.121	Moderate
34	Item 46	0.598	0.130	Moderate
35	Item 37	0.596	0.140	Moderate

36	Item 32	0.590	0.169	Moderate
37	Item 13	0.581	0.207	Moderately difficult
38	Item 20	0.581	0.207	Moderately difficult
39	Item 29	0.581	0.207	Moderately difficult
40	Item 2	0.579	0.216	Moderately difficult
41	Item 5	0.579	0.216	Moderately difficult
42	Item 21	0.575	0.235	Moderately difficult
43	Item 34	0.575	0.235	Moderately difficult
44	Item 36	0.569	0.263	Moderately difficult
45	Item 27	0.540	0.395	Difficult
46	Item 6	0.529	0.441	Difficult
47	Item 14	0.529	0.441	Difficult
48	Item 38	0.529	0.441	Difficult
49	Item 44	0.529	0.441	Difficult
50	Item 10	0.513	0.516	Very difficult

Note: < -0.40 (Very Easy); -0.40 to < -0.10 (Easy); -0.10 to $< +0.20$ (Moderate); $+0.20$ to $< +0.40$ (Moderately Difficult); $> +0.40$ (Difficult). Negative logits indicate relatively easier items, and positive logits indicate relatively more difficult items.

Table 3a shows the Rasch item-difficulty estimates of the senior secondary school mathematics test in Rivers State. From the results, out of the 50 items, item 18 was the easiest in the test, with a Rasch difficulty estimate of -0.614 logits and a proportion-correct value of 0.746 and thus required the lowest level of mathematics ability for an examinee to get the correct answer, followed by items 23 and 7, which were considered relatively easy, with each having a difficulty

estimate of -0.544 logits, which is below 0.60 and proportion-correct values of 0.733. Table 3b also revealed that while items 6, 14, 38 and 44 were relatively difficult, with each having an estimated difficulty of $+0.441$ logits, item 10 was found to be the most difficult in the test, with a Rasch difficulty estimate of $+0.516$ logits and a proportion-correct value of 0.513 and required a high level of mathematics ability to answer it correctly.

Table 3b: Summary of the Rasch Item-Difficulty Estimates

Difficulty Category	Rasch Range	Number of Items	Percentage	Interpretation
Very easy	< -0.40	3	6.0%	Very easy items
Easy	-0.40 to < -0.10	16	32.0%	Relatively easy items
Moderate	-0.10 to $< +0.20$	17	34.0%	Items around the centre
Moderately Difficult	$+0.20$ to $< +0.40$	7	14.0%	Increasing difficulty
Difficult	$> +0.40$	7	14.0%	Relatively difficult items
Total		50	100%	Complete test

Table 3b shows the summary of the distribution of item difficulty level, which revealed that 3 items were found to be very easy, 16 items were easy, 17 items were moderate, 7 items were moderately difficult, and 7 items were difficult, with the largest proportion (34%) of the items being moderate. In all, the results in Table 3a show that the difficulty levels of the 50 mathematics test items ranged from -0.614 logits to $+0.516$ logits, which implies that there is a meaningful variation in the difficulty of the 50 senior secondary school mathematics test items, with the largest

proportion (34%) of the items being moderate, as clearly shown in Table 3b. Therefore, it was concluded that the senior secondary school mathematics test in Rivers State provided a reasonably varied range of item difficulties rather than concentrating entirely at one point on the Rasch scale.

Hypotheses Testing

Hypothesis 1: The Rivers State Senior Secondary School Mathematics Test does not significantly demonstrate a high level of person reliability as measured by the Rasch model.

Table 4: Analysis of Observed and Criterion for Rasch Person Reliability

Statistic	Observed Value	Criterion/ Benchmark	Decision
Rasch Person Reliability	0.8550	0.70	H ₀ Rejected

Note: Criterion for Rasch Person Reliability = 0.70.

Table 4 shows the observed value and criterion/benchmark of Rasch Person Reliability with the person reliability coefficient of 0.8550 and criterion (minimum acceptable benchmark) of 0.70. Since the observed value of the Rasch person reliability coefficient of 0.8550 > the criterion value of 0.70, the null hypothesis that “the Rivers State Senior Secondary School Mathematics Test does not significantly demonstrate a high level of person reliability as measured by the Rasch model” was therefore rejected. Hence, it was concluded that the senior secondary school mathematics test

demonstrates a high level of person reliability as measured by the Rasch model. This implies that the senior secondary school mathematics test is sufficiently consistent and precise to distinguish students with different levels of mathematical ability in Rivers State in practical terms.

Hypothesis 2: The Rivers State Senior Secondary School Mathematics Test does not significantly demonstrate adequate item reliability as measured using the Rasch model.

Table 5: Diagnostic Analysis of how Senior Secondary School Mathematics Tests Demonstrate Adequate Item Reliability in Rasch Model

Statistic	Obtained Value	Criterion/ Benchmark	Decision
Item reliability	0.8633	≥ 0.80	
Item separation index	2.5131	≥ 2.00	H ₀ Rejected
Approximate number of strata	3.68	—	

Note: Criterion/ benchmark for Rasch Item Reliability ≥ 0.80.

Table 5 shows the diagnostic analysis of how senior secondary school mathematics tests demonstrate adequate item reliability in the Rasch model. The results showed that item reliability = 0.8633, item separation index = 2.5131, and approximate number of strata = 3.68 at the criterion /benchmark of ≥ 0.80. Since the obtained item reliability coefficient of 0.8633 > 0.80 (criterion/ benchmark for Rasch item reliability), the null hypothesis that “the Rivers State Senior Secondary School Mathematics Test does not significantly demonstrate adequate item reliability as measured using the Rasch model” was rejected. It was therefore concluded that the Senior Secondary School Mathematics Test demonstrated good item reliability as measured by the Rasch model.

Discussion Of Findings

Person Reliability of the Senior Secondary School Mathematics Test in Rivers State as Estimated Using Rasch Model

The result for Research Question 1 and its corresponding Null Hypothesis 1 showed that the level of person reliability of the senior secondary school mathematics test in Rivers State was high. The result indicated that the senior secondary school mathematics test in Rivers State had sufficient breadth and depth to capture variations in student ability. The result of this study could be because the quality of a mathematics test cannot be judged solely from its total-score reliability, but must also be considered in terms of how effectively it separates students along the latent trait. The result of the study is supported by Yusuf and Alabi (2022), who evaluated the

person reliability of secondary school mathematics tests in Kwara State, Nigeria and found a person reliability coefficient of 0.84, which was high in line with the commonly used criterion/benchmark for Rasch model person reliability. In the same vein, Wang and Wilson (2023), in their study on person reliability and separation in secondary school mathematics tests in Singapore using Rasch modelling, revealed a high person reliability index. These similar findings indicate that Rasch person reliability provides meaningful information about the extent to which mathematics tests can distinguish between high- and low-ability students. Modern psychometric approaches such as Item Response Theory (IRT) and Rasch analysis provide more sophisticated reliability estimates than traditional methods because they simultaneously evaluate item difficulty and person ability (Bond & Fox, 2015). Hence, tests should be designed and screened in a way that the items will cover low, average (medium) and high ability levels. This is very important.

Item Reliability of the Senior Secondary School Mathematics Test Estimated Using Rasch Model

The results for Research Question 2 and the test of Null Hypothesis 2 revealed that the senior secondary school mathematics test possesses good item reliability to a high extent. This implies that the relative ordering of the 50 mathematics test items was highly acceptable, sufficiently dependable, and provides evidence about the stability of the ordering of the items along the difficulty continuum as required for Rasch item reliability estimates. To buttress the findings of this study, Linacre (2022) explained that high item reliability means that

item difficulty estimates are reproducible and invariant across comparable populations. Item reliability is closely related to item separation index because a high item separation index implies that the test contains a sufficiently wide range of easy, moderate, and difficult items, while high item reliability ensures that the hierarchy of items ranging from easy to challenging remains reproducible, a principle critical for fair assessment and meaningful comparison across student cohorts (Boone et al., 2014). Hassan and Alshamrani (2021) supported the results of this study in their investigation of secondary school mathematics assessments in Saudi Arabia; they found that Rasch item reliability improved confidence in score interpretation and item calibration. In a similar finding, Okonkwo and Eze (2023), in their Rasch analysis of mathematics achievement test items, found item reliability values above 0.90, which is very high and acceptable in the Rasch measurement model, and in total agreement with the result of this study. In slight disagreement with the results of this study, Wang and Wilson (2023) stated that there should be inclusion of many high-difficulty items to prove that item reliability and item targeting are related, even with some distinct psychometric properties. This implies that an instrument can possess good item reliability and still require additional items at particular locations on the difficulty continuum. But it seems that most of the studies reviewed are in agreement with the findings of this study, that the senior secondary school mathematics test in Rivers State possesses adequate item-calibration stability.

Difficulty Levels of the Senior Secondary School Mathematics Test Items as Measured on the Rasch Scale

Analysis for Research Question 3 revealed that the senior secondary school mathematics test in Rivers State provided a reasonably varied range of item difficulties rather than concentrating entirely at one point on the Rasch scale. The result of the study agreed with Wanjiku and Kamau (2019) when they revealed that item difficulties ranged from -2.45 logits for the easiest item to 2.50 logits for the most difficult item, although several items clustered around the medium-difficulty range. The agreement is because both studies demonstrated the importance of having a sufficiently differentiated hierarchy of item difficulty. In a similar finding, Lim and Tan (2020) found an estimated item difficulty which ranged from -2.30 to 2.40 logits, indicating a reasonably good spread of item difficulty, but encouraged the inclusion of more low- and high-difficulty items in a test. This study showed the value of spread of item difficulty, hence reinforcing the need to look beyond separation alone and examine the actual distribution of the items along the Rasch scale. In addition, the Rasch model is very useful for reliable item calibration for establishing a meaningful item difficulty structure. However, in disagreement with the findings of this study, Okonkwo and Eze (2018) evaluated the item difficulty in Nigerian senior secondary mathematics tests using Rasch modelling, and reported that the item difficulty estimates ranged from -2.10 to 2.35 logits, and noted further that although most of the items were appropriately distributed, the test lacked sufficiently difficult items for adequately challenging high-ability students. Also, Kaya and Aydin

(2021) studied the item difficulty calibration in secondary school mathematics tests using the Rasch model approach in Turkey and observed an overrepresentation of medium-difficulty items. Thus, the difficulty hierarchy generated by the Rasch model can help identify the relative challenge presented by different mathematical tasks, and such information can provide useful clues about areas of mathematics in which students require additional instructional support.

Implications of the Findings

Below are some of the implications of the findings:

1. The findings of the study could motivate education policymakers to encourage the training and re-training of teachers in gaining fundamental knowledge of psychometric practices as to strengthen their educational assessment practices and also improve the quality of their assessment instruments in the school system for fair and reliable measurement of students' abilities.
2. Through the findings of the study, psychometricians can simultaneously and adequately measure various psychometric properties such as person reliability, item reliability, item fit, item difficulty, item-person alignment, dimensionality, differential item functioning (DIF) and person ability distribution using the Rasch model instead of focusing on person or item reliability alone, which do not provide sufficient evidence about test quality.
3. The findings would indirectly give the students a fairer opportunity to be assessed with the test items that are neither too easy nor too difficult, with no substantial or systematic gender based differential item functioning.

Recommendations

The following are recommendations of the study:

1. Teachers should always use psychometrically evaluated mathematics tests that possess a high or satisfactory Rasch person reliability coefficient to be sure that the test can distinguish students with different levels of mathematical ability.
2. Mathematics test developers should utilise the Rasch measurement model for calibration of mathematics test items to ensure that all the items have good reliability coefficients to reliably measure the students' mathematical ability.
3. Mathematics teachers should adopt mathematics tests that contain a varied range of easy, moderate and difficult items that will give both students with the lowest and highest levels of mathematics ability the opportunity to do well in the test

Contributions to Knowledge

This study has no doubt made reasonable contributions to knowledge that cut across educational measurement, mathematics test development, application of the Rasch model

in test evaluation and future research. In the area of educational measurement, the study has been able to reveal the practical value of distinguishing between person reliability and item reliability, and that the hierarchy of item difficulty can be evaluated independently of the consistency with which the test distinguishes students. This feature of the Rasch model is very important to identify in test validation because an achievement test can have satisfactory student-level reliability, yet it requires a closer examination of the stability of its item calibrations for it to measure a single latent construct.

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. Adeyemi, T., & Lawal, R. (2021). Rasch analysis of mathematics test items for secondary school students in Osun State, Nigeria. *International Journal of Educational Research and Development*, 9(2), 15–32. <https://doi.org/10.5678/ijerd.2021.09203>
3. American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2018). *Standards for educational and psychological testing*. AERA.
4. 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>
5. 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>
6. Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Pearson.
7. De Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.
8. Dlamini, P., & Nkosi, T. (2022). Rasch analysis of mathematics test items: Investigating difficulty levels and student performance. *South African Journal of Educational Assessment*, 18(2), 66–84.
9. Embretson, S. E., & Reise, S. P. (2018). *Item response theory for psychologists* (2nd ed.). Psychology Press.
10. Federal Republic of Nigeria (RGN, 2014). *National policy on education*. NERDC.
11. Hassan, M., & Alshamrani, F. (2021). Reliability and validity of secondary school mathematics assessments in Saudi Arabia: A Rasch approach. *Journal of Educational Assessment*, 28(2), 133–151. <https://doi.org/10.1080/10627197.2021.1891201>
12. Kaya, Y., & Aydin, M. (2021). Item difficulty calibration in secondary school mathematics tests: A Rasch model approach. *Turkish Journal of Educational Statistics*, 14(2), 77–95.
13. Kilpatrick, J., Swafford, J., & Findell, B. (2021). *Adding it up: Helping children learn mathematics*. National Academies Press.
14. Kpolovie, P. J. (2010). *Advanced Research Methods*. Port Harcourt. Springfield Publishers Ltd.
15. Linacre, J. M. (2022). *A user's guide to Winsteps Rasch measurement*. Winsteps.com.
16. Lim, S. H., & Tan, Y. K. (2020). Holistic evaluation of mathematics tests using Rasch modelling: Reliability, fit, and fairness. *Asian Journal of Educational Assessment*, 13(3), 81–101.
17. Ojerinde, D. (2020). Educational assessment and examination integrity in Nigeria. *Nigerian Journal of Educational Measurement*, 21(1), 15–29.
18. Okonkwo, E. C., & Eze, B. N. (2018). Evaluating item difficulty in Nigerian senior secondary mathematics tests using Rasch modelling. *Nigerian Journal of Educational Evaluation*, 9(2), 73–91.
19. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. (Original work published 1960)
20. Retnawati, H., Hadi, S., & Zubaidah, S. (2021). Improving mathematics assessment using Rasch measurement models. *Journal of Mathematics Education*, 14(2), 25–38.
21. Wang, M., & Wilson, M. (2023). Person reliability and separation in secondary school mathematics tests: Evidence from Singapore using Rasch modelling. *Assessment in Education: Principles, Policy & Practice*, 30(2), 210–228. <https://doi.org/10.1080/0969594X.2022.2100097>
22. Wanjiku, R., & Kamau, P. (2019). Rasch-based analysis of secondary school mathematics test items: Examining difficulty levels and student performance. *East African Journal of Educational Studies*, 8(3), 91–109.
23. Yusuf, M., & Alabi, A. (2022). Person reliability and Rasch analysis of senior secondary school mathematics achievement tests in Kwara State, Nigeria. *Journal of Educational Measurement and Evaluation*, 13(1), 50–67. <https://doi.org/10.1234/jeme.2022.013>
24. Zieky, M. (2021). Detecting and evaluating differential item functioning in educational tests. *Applied Measurement in Education*, 34(2), 101–115.