International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Assessment of the Psychometric Properties of the Delta State Basic Education Certificate Mathematics Examination Using Item Response Theory

Ejegreh, O. R.[1], *Osadebe P. U.*[2], *Ossai P. A. U.*[3]

[1]Department of Guidance and Counselling, Delta State University Abraka, Nigeria
[2]Department of Guidance and Counselling, Delta State University Abraka, Nigeria, osadebepu@delsu.edu.ng
[3]Department of Guidance and Counselling, Delta State University Abraka, Nigeria, ossaipeter@delsu.edu.ng

ABSTRACT

The study assessed the psychometric properties of the Delta State Basic Education Certificate mathematics examination using item response theory. Three research questions guided the study. The multiple triangulation research design was used. The population comprised 140,959 Junior Secondary School (JSS) 3 students in the 2022/2023 academic session in Delta State. A sample of 1,000 students was selected through a multistage sampling procedure. The main instrument that was used for the study is the Mathematics Achievement Test (MAT) used in the Delta State Basic and Education Certificate in the 2022. Pearson separation reliability index and Item Characteristics Curve were used to answer research question 1; Chi-square goodness of fit was used to answer research question 2; while factor analysis using Principal Component Analysis of the varimax method was used to answer research question 3. The findings of the study revealed that the reliability of the MAT instrument is high when compared within the framework of IRT as indicated by the value is 0.70; that majority of the items in the test (55 out of 60) have a good fit in the overall model while 5 did not have a good fit; and that majority of the items in the test measured a single construct, as shown in the scree plot. The study recommended amongst others, that examination bodies responsible for test development should consider revising the test to ensure a more balanced distribution of difficulty levels. This can help mitigate potential biases and ensure that the test effectively assesses a wide range of abilities.

Keywords: Assessment; Psychometric Properties; Delta State Basic Education Certificate Mathematics Examination; Item Response Theory.

Introduction

The problem JSS 3 students face during their mathematics examination poses a serious issue that needs urgent attention. Some may put the blame on the content students are required to learn, while others blame the instructional method taken by teachers. However, it is very glaring that the issue has an immense effect on students' achievement in mathematics both in internal and external examinations, which have caused a decline in the percentage of passes in mathematics (Akinsola & Ayodele, 2021; Onwuachu et al., 2023). Every year the number of students who write the 'resit' examination in the subject increases despite the fact that the subject is taught every day in school (Emaikwu, 2022).

Observations over the years have not only indicated that JSS 3 students perform poorly in mathematics but also shown that the errors students commit while solving mathematics problems have contributed to their poor performance in mathematics, leading to many failures even after writing the 'resit' examination (Udo & Ibe, 2020; Nwaubani & Okwudishu, 2022). Recent mathematics education research indicated that readiness tests were developed and used to determine the preparedness levels of students advancing from primary six to Junior Secondary School One (JSS 1), where they begin a new mathematics programme, and similarly for students transitioning to senior secondary mathematics (Okafor & Olatunji, 2021; Adegboye et al., 2024). With the emphasis placed on the teaching and learning of mathematics, as well as the usefulness of the subject and the good condition of service of Delta State teachers, it is expected that no fewer than 90% of the students should perform above average in the JSSCE mathematics. However, this has not been the case (Olatoye & Ajayi, 2023).

With the persistent failure among JSS 3 students in mathematics examinations, one therefore needs to ask whether the challenges students face during their mathematics examination are a result of test items or their readiness. Are the items reliable? Do the items fit the ability levels of the students? Do the items meet the unidimensionality assumption of Item Response Theory (IRT)? Based on these questions, this study aims to assess the psychometric properties of the Delta State Basic Education Certificate Mathematics Examination using the Item Response Theory model.

This study is anchored on the Item Response Theory (IRT). Three of the pioneers who pursued parallel research independently were the Educational Testing Services psychometricians known as Frederic M. Lord, the Danish mathematician, George Rasch, and Austrian sociologist Paul Lazarsfeld (Millikarjuna, 2014). Lord (1952) brought the idea of latent trait or ability and at the same time differentiated this construct from observed score.

Lazarsfeld (1950) only described the unobserved variable as accounting for the observed interrelationships among the item responses. While Rasch (1960) reported the need for creating statistical models that maintain the property of specific objectivity, the idea that people and item parameters be estimated separately but comparable on similar metric.

Item response theory (IRT) rests on two basic postulates: (a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increase (Jessa, et al., 2023).

In using IRT Model, it is very important to assess the extent to which the IRT model assumptions are valid for the given data. The most significant assumptions common to all IRT models is unidimensionality, other assumption relates to IRT is Local independence. The test data can only be valid for latent trait model estimation only if these assumptions are met. The assumption of unidimensionality assumes that a set of items and/or a test measure(s) only one latent trait (Kyung, 2013). This implies that the performance of each examinee is assumed to be governed by a single factor, referred to as ability. Since individuals' cognitive and personal characteristics, influence test performance and cannot often be controlled; it is not always possible to meet this assumption. One can then talk about the unidimensionality of a test only when there is just one dominant ability in it (Hambleton, et al., as cited in Orheruata, 2015).

To satisfy this assumption, one can apply any of these eleven methods for testing for unidimensionality as cited by Ojerinde and Ifewulu (2012); Cronbach analysis test, Factor analysis, Eigenvalue test, Random baseline test, Biserial test, Factor loading test, Congruence test, Part/Whole test, Communality test, Vector frequency test and Confirmatory factor analysis (F.A) and Structural equation modelling (SEM) test, using the SPSS package. A support for the unidimensionality of the items in the scale is provided when the model fits the data well and there are no noteworthy residual correlations (i.e., no such correlations greater than or equal to 0.20) (Ojerinde, 2013). Any violation of this assumption would result in inadequacy of the model in describing the data and hence unreliable estimation of the examinee's ability. Therefore, the correct specification of the number of the latent dimensions is directly tied to the construct validity of the test (Rijn, et al., 2016).

Local independence refers to the assumption that there is no statistical relationship between examinees' responses to the pairs of items in a test, once the primary trait measured by the test is removed (Kyung, 2013). It implies that the probability of an examinee getting an item correct is unaffected by the answer given to other items in the test. i.e. questions should be set so that no one item gives an insight to the answer of the other. Local independence according to Ojerinde (2013) does not mean that items do not correlate with each other, but that performance on different items is independent but conditional on the student's ability. Thus, the probability that a student will answer correctly any two items must be the product of the probability that the student will answer correctly each separate item. Also, the association between two items should not differ significantly from zero, otherwise, it may be said that the responses to the items are influenced by other extraneous factors other than what the instrument is designed to measure (Ojerinde, 2013).

The theory was adopted due to its applicability to sift through item level statistics. IRT generated new rules of measurement and presented as modern and superior alternative to CTT (De Boeck & Wilson, 2004; Embretson & Reise, 2010; Nering & Ostini; 2010, Zickar & Broadfoot, 2008). In using IRT, one can assess through the item characteristics within a multiple item test and estimate the examinee's ability given the item parameters and the response pattern to the test by that examinee. A more comprehensive approach to psychometrics that rectifies many of the perceived shortcomings associated with classical approaches is also provided by IRT. It employs new concepts to describe tests (item characteristics curve, item/test information function and so on). It puts the focus on the estimation of item's operational characteristics like assessment of test dimensionality, estimating of the difficulty, discriminating, guessing parameters, item bias and differential item functioning. IRT methods differentiate error more finely, most especially with respect to characteristics of individual items that may affect their performance. A goal of IRT is to enable a researcher to establish certain characteristics of items that are independent of who is completing them. It examines the level of the attribute being measured that most strongly influences an item. The purpose of IRT is to estimate both the value of the latent trait for each respondent and the item parameters for each item. In IRT, each item specifies three parameters that define an s-shape logistic curve, called item characteristic curve (ICC), linking probabilities of individuals into position of the individuals in the latent trait (or ability).

Research Questions

The following research questions were raised to guide the study:

1. What is the reliability index of the Delta State Basic Education Certificate Mathematics Examination?
2. What is the overall model fit of the Delta State Basic Education Certificate Mathematics Examination?
3. What is the evidence of unidimensionality of the Delta State Basic Education Certificate Mathematics Examination?

Methods

The multiple triangulation research design was used for the study. Kpolovie (2014) asserts that this design enables a multi-method approach to investigating an instrument's psychometric properties. This approach is suitable since the study's goal was to assess the psychometric properties of the Delta State Basic Education Certificate Mathematics Examination using IRT model. The population of this study comprised 140,959 Junior Secondary

School (JSS) 3 students in the 2022/2023 academic session in Delta State. A sample of 1,000 JSS 3 students was used for the study through systematic sampling and multistage sampling procedure to select schools across the various Local Government Areas that make up Delta State. In the first stage, the researcher used proportional stratified sampling technique. The proportional stratified sampling technique is a probability sampling approach in which the different strata are identified and the number of elements drawn from each stratum is proportionate to the relative number of elements in each stratum. The choice of this sampling technique is because there were not an equal number of students across all of the Local Governments Areas that were included in the study. Therefore, it was necessary to depict the student population in each local government area. In doing so, the researcher calculated the sample size's proportion to the population as a whole, arriving at 0.71%. In the second stage, the researcher utilized a simple random sampling technique to choose one school from each of Delta State's 25 Local Government Areas as the schools to be used. The researcher did this by writing the names of all the schools in each local government on a piece of paper, folded it, and then poured the contents onto a container. She then chose one paper and unfolded it to show what was written on it. Schools chosen through this procedure were selected. In the third stage, the researcher used a convenience sampling technique to choose the students who participated in each school. The researcher chose only available students who were willing to engage in the study and met the requirement of being a JSS 3 students at the chosen school when employing the convenience sampling technique.

The main instrument that was used for the study is the Mathematics Achievement Test (MAT) used in the Delta State Basic and Education Certificate in the 2022, which was obtained from the Exam and Standard in Delta State. The test contains 60 multiple choice questions having four options; one key and three distracters. In ascertaining the validity of the instrument, the researcher ensured that the test items have content, construct and face validity. Content validity of the test is the extent to which a given content areas are covered, the areas taught as in the case of achievement or component areas not taught as the case of aptitude. A table of specification was used to establish high content validity (Ukwuije & Opara, 2012). In order to establish the content validity of the instrument, a table of specification was drawn. This showed the various sections of the achievement test that were considered and also the total number of the items which were included in the test. Face validity of the test was determined through the inspection of the items by experts in Mathematics to ensure that the test items look like what is being measured. The face validity of the instrument was ascertained by the comments and suggestions of experts in the area of Mathematics and Measurement and Evaluation. Construct validity of test was determined when the psychological construct of achievement is used for a test. The tests were compared with sections of existing test in the subject area and they were closely related. Experts in psychological testing and Mathematics as well as the project supervisor verified the construct validity of the test. To establish the reliability of the test items, Kuder-Richardson formula 20 was used. The Kuder-Richardson formula 20 was applied for true/false or dichotomously scored data where the responses are scored either right or wrong, pass or fail. In the binary–type of data, the correct answer or option was treated as one while the incorrect or wrong option was treated as zero.

The instrument was administered to the JSS 3 students in the sampled schools by the researcher and five research assistants. Guidelines were given to the research assistants to ensure conformity in test administration across the sampled schools. The students were required to indicate the correct option to each question by circling the correct option. The sample for item analysis was treated respectively. Pearson separation reliability index and Item Characteristics Curve were used to answer research question 1; Chi-square goodness of fit was used to answer research question 2; while factor analysis using Principal Component Analysis of the varimax method was used to answer research question 3.

Results

Research Question 1: What is the reliability index of the Delta State Basic Education Certificate Mathematics Examination?

Table 1: Summary Statistics for the Total Scores indicating person separation reliability of the MAT

Test	Items	Alpha	Mean	SD	Skew	Min	Median	Max	IQR
Full Test	60	0.71	21.23	5.97	0.53	8.00	20.00	38.00	8.00

As shown in Table 1, the Alpha value is 0.71 which tends towards 1.00, indicating a strong reliability. Therefore, the reliability of the MAT instrument is high when compared within the framework of IRT as indicated by the Alpha value is 0.70.

Research Question 2: What is the overall model fit of the Delta State Basic Education Certificate Mathematics Examination?

Table 2: Item Fit Statistics of the MAT

S/N	Item	$S\text{-}X^2$	Sig.	Remark
1.	MAT7	42.8491	0.532	√
2.	MAT14	45.2966	0.512	√
3.	MAT15	48.2094	0.499	√
4.	MAT37	60.0640	0.448	√
5.	MAT22	62.8076	0.431	√
6.	MAT1	100.4180	0.43	√

S/N	Item	*S-X*^{2}	Sig.	Remark
7.	MAT2	76.8868	0.423	√
8.	MAT3	69.0270	0.406	√
9.	MAT4	133.0707	0.386	√
10.	MAT5	107.5936	0.367	√
11.	MAT6	112.1945	0.339	√
12.	MAT1	117.3271	0.33	√
13.	MAT8	107.3969	0.329	√
14.	MAT59	126.1015	0.329	√
15.	MAT58	175.3953	0.327	√
16.	MAT56	158.3176	0.322	√
17.	MAT57	126.4762	0.322	√
18.	MAT54	177.7030	0.32	√
19.	MAT55	83.4003	0.32	√
20.	MAT9	121.8046	0.317	√
21.	MAT53	117.5303	0.309	√
22.	MAT52	154.6540	0.304	√
23.	MAT10	162.4243	0.299	√
24.	MAT11	155.8290	0.295	√
25.	MAT51	127.7897	0.293	√
26.	MAT50	140.6044	0.282	√
27.	MAT49	205.2677	0.271	√
28.	MAT48	222.0791	0.27	√
29.	MAT12	155.0588	0.265	√
30.	MAT47	121.3129	0.265	√
31.	MAT13	168.1758	0.264	√
32.	MAT46	116.9866	0.257	√
33.	MAT45	111.2053	0.256	√
34.	MAT44	126.4468	0.243	√
35.	MAT43	97.2216	0.241	√
36.	MAT16	164.5158	0.236	√
37.	MAT17	99.6796	0.224	√
38.	MAT42	97.5991	0.22	√
39.	MAT18	164.7874	0.217	√
40.	MAT41	88.7719	0.212	√
41.	MAT40	134.8054	0.193	√
42.	MAT39	168.1052	0.183	√
43.	MAT19	148.7843	0.182	√

S/N	Item	*S-X*^{2}	Sig.	Remark
44.	MAT38	87.1004	0.166	√
45.	MAT36	145.6993	0.146	√
46.	MAT35	193.8926	0.14	√
47.	MAT34	126.8538	0.136	√
48.	MAT33	85.4975	0.129	√
49.	MAT32	118.7728	0.125	√
50.	MAT31	439.7119	0.12	√
51.	MAT30	143.4017	0.096	√
52.	MAT20	151.8127	0.092	√
53.	MAT21	114.7221	0.089	√
54.	MAT29	82.0046	0.08	√
55.	MAT28	80.1091	0.057	√
56.	MAT27	199.1777	0.047	X
57.	MAT23	209.3166	0.041	X
58.	MAT24	120.1391	0.028	X
59.	MAT25	173.8338	0.026	X
60.	MAT26	174.7059	0.008	X

Key: √ = Good Fit; X = Not Good Fit

Criterion = p>0.05

As shown in Table 2, the p-value ranged from 0.008 to 0.532. Items with p-value greater than 0.05 are regarded as having a good fit in the overall model while items with p-value less than 0.05 are regarded as having no good fit in the overall model. Based on this criterion, out of a total of 60 items, 55 have a good fit in the overall model while 5 did not have a good fit.

Research Question 3: What is the evidence of unidimensionality of the Delta State Basic Education Certificate Mathematics Examination?

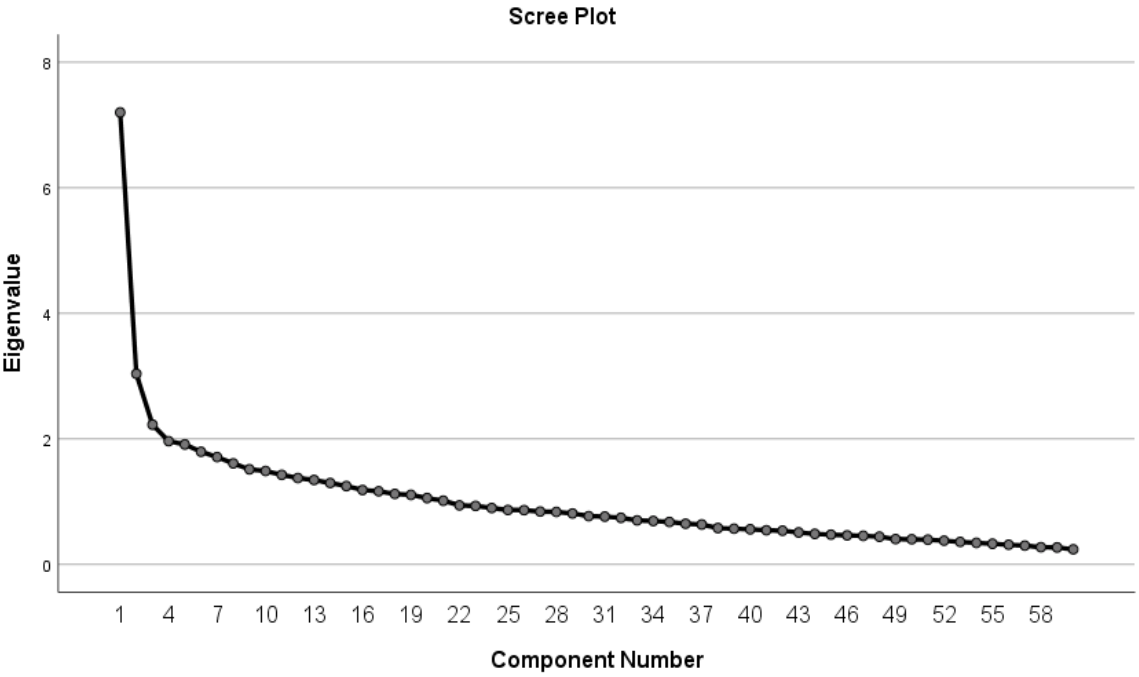

Figure 1: Scree Plot for the MAT

Figure 1 shows the scree plot for the MAT. From the figure, a careful examination of the scree plot shows that there is only one construct before the breaking point or elbow joint. This therefore succinctly shows the unidimensionality of the underlining construct of the MAT. All the items measure one construct.

Discussions

The first finding showed that the reliability of the MAT instrument is high when compared within the framework of IRT as indicated by the value is 0.70. In the context of IRT, reliability refers to the consistency and stability of test scores across different administrations and among different groups of test-takers. An Alpha value of 0.70 indicates that the test items collectively contribute to a reliable assessment of academic talent, with minimal measurement error. This means that the MAT instrument can consistently distinguish between individuals with varying levels of academic talent, providing dependable results that can be trusted for decision-making purposes. The high reliability of the MAT instrument within the framework of IRT is particularly noteworthy because IRT takes into account the characteristics of both the test items and the test-takers when estimating reliability. By considering the difficulty and discriminatory power of each item, as well as the ability level of the test-takers, IRT offers a more nuanced and accurate assessment of reliability compared to classical test theory approaches. A reliability coefficient of 0.70 is generally considered acceptable for many educational and psychological assessments, indicating that the MAT instrument is well-suited for its intended purpose of identifying and measuring academic talent. However, it's essential to interpret this value in conjunction with other validity evidence and consider the specific context and use of the assessment. This finding is in line with Orangi and Dorani (2010), who carried out research to develop a social studies aptitude test for high school students based on item-response theory (IRT), and found that the constructed forms were of high reliability, they were at the same time acknowledgeable through the analysis based on the classical Method and they were also in accordance with the three–factors of the Item Response Theory. The finding also agrees with Ezechukwu, Chinecherem, Oguguo, Ene and Ugorji (2020), who determined the psychometric properties of the Economics Achievement Test (EAT) using Item Response Theory (IRT). Two popular IRT models namely, one-parameter logistics (1PL) and two-parameter logistics (2PL) models were utilized just like in the current study. Reliability and validity for each item and for the whole test were established according to the one-parameter and two-parameter logistic models. The finding revealed that the instrument was highly reliable and fit for use.

The second finding revealed that majority of the items in the test (55 out of 60) have a good fit in the overall model while 5 did not have a good fit. This finding underscores the robustness and effectiveness of the test construction process. A good fit within the model indicates that these items align well with the underlying theoretical framework or measurement model used to guide the test development. This alignment enhances the validity and reliability of the test, ensuring that it accurately assesses the intended construct or domain. Items that demonstrate a good fit within the model contribute to the coherence and consistency of the assessment, providing meaningful and interpretable results. Their inclusion enhances the discriminative power of the test, allowing it to effectively differentiate between individuals with varying levels of proficiency or ability in the target domain. However, the finding that five items did not exhibit a good fit within the overall model warrants attention and further investigation. Items that deviate from the expected pattern or fail to align with the underlying model may introduce noise or error into the assessment process, potentially compromising the validity and reliability of the test results. The identification of these misfitting items presents an opportunity for refinement and improvement in the test construction process. By carefully scrutinizing the problematic items and exploring potential sources of misfit, test developers can enhance the quality and effectiveness of the assessment. This may involve revising or reevaluating the content, format, or scoring of the misfitting items to better align them with the intended measurement model. This finding agrees with Thompson (2019), who found that well-fitting items ensure their difficulty levels and discrimination power are accurately measured, leading to reliable and valid score interpretations and that items with poor fit can distort score estimates and compromise the overall quality of the test. The findings are also consistent with Oku and Iweka (2018), who used the Maximum Likelihood Estimation Method to perform item analysis on each item and discovered that 99 of the test items fit the One-Parameter Model (1-PLM). The findings back up the findings of Ani (2014), who used the maximum likelihood estimate approach of BILOG-MG computer programming to analyse the data and discovered that all 50 Economics test items passed item calibration.

The third finding showed that majority of the items in the test measured a single construct, as shown in the scree plot. The model assumes that there is one dominant latent trait being measured by the test and that this trait is the driving force for the responses observed for each item in the Mathematics Achievement Test. This finding is significant in validating the underlying theoretical framework guiding the assessment. The scree plot, often used in factor analysis, provides insight into the number of latent factors or constructs underlying the observed responses to test items. In this case, the scree plot suggests that there is one dominant latent trait driving the responses observed for each item in the Mathematics Achievement Test (MAT). The assumption that there is a single dominant latent trait being measured by the test aligns with the theoretical framework guiding the development of the MAT. This framework posits that mathematical achievement is a multidimensional construct that can be effectively captured and assessed by a set of items targeting a common underlying trait. By focusing on a single dominant trait, the MAT aims to provide a comprehensive and reliable measure of individuals' mathematical proficiency. The identification of a single dominant construct through the scree plot reinforces the coherence and validity of the test design. It suggests that the items included in the MAT are indeed tapping into a common underlying dimension of mathematical achievement, rather than measuring unrelated or overlapping constructs. This clarity in construct measurement enhances the interpretability and utility of the test results, allowing educators, researchers, and policymakers to make informed decisions based on individuals' performance. Moreover, the acknowledgment of a single dominant trait in the MAT underscores the importance of item selection and construction in ensuring that the test effectively captures the intended construct. Items that align closely with this dominant trait contribute most strongly to the overall measurement accuracy of the test, whereas items that deviate may introduce noise or error into the assessment process. Several researchers have used factor analysis to determine the unidimensionality of a test and were successful. For instance, Kpolovie and Emekene (2016) validated the advanced progressive matrices for Nigerian sample using Item Response Theory. They used factor analysis to determine the unidimensionality of the scale and found that the unidimensionality of the underlining construct of the APM scale, namely intelligence or fluid ability and that all 36 items of the scale measure one construct, the fluid ability of the test taker as confirmed by the scree plot. They concluded that all the items APM unquestionably measure just one general intelligence factor in Nigeria just as it does in all other countries that the test is actively in use.

Conclusion and Recommendations

Based on the findings, it can be concluded that the high reliability of the MAT instrument, assessed within the framework of Item Response Theory (IRT), reinforces confidence in its ability to consistently measure the intended constructs. Additionally, while most items fit well within the overall model, a few exhibits a less-than-ideal fit, suggesting potential areas for refinement. The scree plot analysis indicates that the majority of items effectively measure a single construct, contributing to the test's validity and coherence. Based on the findings from this study the following recommendations were made:

1. While the MAT instrument demonstrates high reliability overall, ongoing monitoring and evaluation are essential to maintain this level of consistency.
2. Government should investigate the factors contributing to the poor fit of certain items within the overall model.
3. Examination bodies should validate the assumption that most items measure a single construct by conducting further analyses, such as confirmatory factor analysis or multitrait-multimethod modelling.

References

Adegboye, A., Olorunfemi, T., & Ige, A. (2024). Transition challenges in mathematics learning: Readiness of Nigerian students from primary to secondary schools. *Journal of Mathematics Education Research, 18*(1), 55–70.

Akinsola, M. K., & Ayodele, J. B. (2021). Pedagogical strategies and students' mathematics performance in Nigerian secondary schools. *African Journal of Educational Studies, 17*(2), 89–104.

De-Boeck, P. & Wilson, M. (2004). *Explanatory item response model: a generalized linear e nonlinear approach.* New York: Springer.

Emaikwu, S. O. (2022). The increasing trend of resit examinations in mathematics among Nigerian secondary school students. *International Journal of Educational Assessment, 9*(3), 33–47.

Embretson, S. E., & Reise, S. P. (2010). Multivariate applications books series. *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Idowu, O., & Afolabi, A. (2020). Item response theory analysis of mathematics multiple-choice items in Nigerian junior secondary certificate examinations. *International Journal of Testing and Measurement, 12*(4), 211–229.

Jessa, M. O., Odili, J. N., & Osadebe, P. U. (2023). Development of Social Studies Aptitude Test for Testing Critical Thinking Skills: Implication for the Achievement of Education for Sustainable Development (ESD). *Canadian Journal of Educational and Social Studies*, 3(4), 99-119. DOI: 10.53103/cjess.v3i4.163

Kpolovie, P. J. (2014). *Test, measurement and evaluation in education* (2nd edition). New Owerri: Spring Field Publishers Ltd.

Kyung, T. H. (2013). *Windows software that generates IRT parameters and item responses: Research and evaluation program methods (REMP).* University of Massachusetts Amherst. Retrieved from https://www.umass.edu/remp/software/simcata/wingen/homeF.html

Lazarsfeld, P. F. (1950). *The logical and mathematical foundation of late4nt structure analysis. In S. A. Stouffer, L. Guttman, Measurement and prediction.* New York: Wiley. 362-412.

Lord, F. M. (1952). *Application of item Response theory to practical testing Problems.* New Jersey; Lawrence Erlbaum Association, Inc.

Nwaubani, O. O., & Okwudishu, C. (2022). Error patterns and their influence on students' achievement in junior secondary mathematics. *Journal of Contemporary Issues in Education, 10*(2), 101–115.

Oguguo, B. C. E., & Lotobi, R. A. (2019). Parameters of Basic Science Test Item's of 2011 Basic Education Certificate Examination Using Item Response Theory (IRT) Approach in Delta State, Nigeria. *European Journal of Educational Sciences, 6*(1), 22-36.

Ojerinde, D. (2013). Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results. Retrieved from https://ui.edu.ng/sites/.../PROF%20OJERINDE'S%20LECTURE%20(Autosaved).pdf

Ojerinde, D., & Ifewulu, C. (2012). Basic concepts and principles of psychometrics. Ibadan: Stirling-Horden Publishers.

Okafor, C., & Olatunji, A. (2021). Assessing mathematics readiness for junior secondary school students in Nigeria. *Journal of Educational Research and Development, 15*(1), 73–86.

Olatoye, R. A., & Ajayi, O. (2023). Mathematics education and student achievement in Nigeria: Implications for curriculum reform. *Nigerian Journal of Educational Evaluation, 15*(2), 65–80.

Onwuachu, C., Nwosu, P., & Igwe, U. (2023). Instructional delivery and mathematics achievement of secondary school students in Nigeria. *Journal of Mathematics and Science Education, 12*(1), 44–58.

Orheruata, M. U. (2015). Item Parameter Drift in Certificate Examinations and it's implication on Decision Making. *African Journal of Theory and Practice of Educational assessment. Educational Assessment*, 2, 98-105.

Ostini, R., & Nering, M. (2010). *Polytomous item response theory models: Quantitative Applications in the Social Sciences,* 144. Thousand Oaks, CA: Sage Publications.

Rasch, G. (1960). *Pobabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

Rijn, R. W. V., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey. *Assessments.* DOI: 10.1186/s40536-016-0025-3.

Suleiman, M., & Mohammed, A. (2022). Psychometric evaluation of mathematics examination items using item response theory in Nigerian secondary schools. *Journal of Educational Measurement and Statistics, 14*(3), 145–162.

Thompson, B. (2019). *Item analysis with the Rasch model.* ERIC.

Udo, E., & Ibe, F. (2020). Analysis of students' errors in mathematics and implications for teaching in junior secondary schools. *Journal of Educational Review and Practice, 21*(2), 56–70.

Zickar, M. J. & Broadfoot, A. A. (2008). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C.E. Lance & R. J. Vandenberg (Eds.) *Statistical and methodological myths and urban legends*. Routledge Academic.