ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2026; 5; 55-81

Assessing the Dimensionality and Model Fit of the 2015 NECO Mathematics Multiple-Choice Items in Nigeria Using the Four-Parameter Logistic Model for Achieving SDG 4

[a]Opesemowo Oluwaseyi Aina Gbolade*, [b]Opatunji Kehinde Olufunke, [c]Babatimehin Temitope & [d]Opesemowo Titilope Rachael

[ad]University of Johannesburg, South Africa
[bc]Obafemi Awolowo University, Ile-Ife, Nigeria

Abstract

Evidence on the psychometric quality of large-scale examinations remains essential for ensuring valid educational decisions and monitoring learning outcomes. This study examined the dimensionality and model-fit of the 2015 National Examinations Council (NECO) Mathematics items in Nigeria using Four-Parameter Logistic (4PL) item response theory model. An ex-post-facto research design was adopted. The population comprised 24,718 candidates who wrote the 2015 NECO Senior School Certificate Mathematics III Examination in Osun State, Nigeria. Using a multistage sampling procedure, responses from 3,000 examinees were selected for analysis. Data were analysed using the MIRT package in R, with item parameters estimated through marginal maximum likelihood estimation and the expectation-maximisation algorithm. Dimensionality was assessed using factor analytic procedures, while model-fit was evaluated using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted Bayesian Information Criterion (SABIC), and Hannan-Quinn Information Criterion (HQ). Results indicated that the unidimensional model provided a superior fit to the data compared with multidimensional alternatives, yielding lower information criterion values (AIC=1,142,691; SABIC=1,144,092; HQ=1,145,043; BIC=1,145,043). Item parameter estimates further indicated that most items demonstrated acceptable discrimination, moderate difficulty, low pseudo-guessing parameters, and minimal upper-asymptote effects. These results indicate that the examination items functioned satisfactorily within the 4PL framework and met the unidimensionality assumption. The study provides empirical evidence for the psychometric soundness of the 2015 NECO Mathematics examination and underscores the importance of rigorous measurement practices to enhance the validity, fairness, and accountability of large-scale educational assessments in support of Sustainable Development Goal 4 (SDG 4).

Keywords: Item Response Theory (IRT), Four-Parameter Logistic Model (4PL), Model fit, Dimensionality, Sustainable Development Goal 4 (SDG 4).

University of Johannesburg, South Africa, e-mail: oopesemowo@uj.ac.za

To cite this article:

Gbolade, O. O. A., Olufunke, O. K., Temitope, B., & Rachael, O. T. (2026). Assessing the dimensionality and model fit of the 2015 NECO mathematics multiple-choice items in Nigeria using the four-parameter logistic model for achieving SDG 4. *Journal of Computerized Adaptive Testing in Africa, 4*, 55-81. https://doi.org/10.71291/6rtdfq13

Received: 13 April 2026
Accepted: 30 June 2026
Published: 29 July 2026

INTRODUCTION

Education remains a fundamental catalyst for achieving Sustainable Development Goal 4 (SDG 4), which seeks to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all (Fuentes-Penna et al., 2025). A key indicator of quality education is the validity and reliability of assessment systems that measure learners' knowledge, skills, and competencies. Examinations are central to this process, as systematic tools for determining how students have acquired the intended learning outcomes (Goel et al., 2021). In Nigeria, the academic progress and transition of secondary school students to higher education are primarily determined by their performance in national standardised examinations, such as those conducted by the National Examinations Council (NECO), the West African Examinations Council (WAEC), and the National Business and Technical Examinations Board (NABTEB).

The NECO Senior School Certificate Examination (SSCE), particularly in Mathematics, is a crucial benchmark for assessing learners' proficiency in logical reasoning, problem-solving, and quantitative skills, which are essential for scientific and technological advancement. Mathematics is a compulsory subject across educational levels and a prerequisite for admission into tertiary institutions, making it a core determinant of academic mobility and employability. The Nigerian Educational Research and Development Council (NERDC, 2013) structured the secondary school Mathematics curriculum around the domains of algebraic processes, number and numeration, geometry, and statistics, which are integral to everyday reasoning and technological literacy. Despite its importance, persistent reports of poor performance among Nigerian students in Mathematics (Ogunbanwo, 2014; Ohanyelu, 2021) raise concerns not only about instructional quality but also about the psychometric integrity and fairness of national assessment systems.

Assessment quality hinges on the soundness of the test items used to evaluate learners' abilities. Test items must accurately represent the measured construct, differentiate between ability levels, and produce consistent results across populations. Within the broader SDG 4 framework, valid and equitable assessment systems are essential because educational policy decisions, certification processes, and access to further learning opportunities depend heavily on the accuracy of measurement outcomes. Consequently, weaknesses in test construction or model misfit may lead to inaccurate learner classification, inequitable educational decisions, and distorted interpretations of educational quality. In psychometric research, two theoretical frameworks are traditionally employed to evaluate item quality: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT assumes that a test score comprises a true score and an error component ($X = T + E$), focusing primarily on overall test reliability rather than item-level analysis (Wang, 2017). Although simple and widely applied, CTT's estimates of item difficulty and discrimination are sample-dependent, which limits generalizability and precision, especially at the extremes of the ability distribution (Hussein & Gasmalla, 2023).

To overcome these limitations, IRT has emerged as a more robust approach to modelling the relationship between examinees' latent abilities and their probability of responding correctly to individual items. IRT treats ability and item characteristics as independent parameters, allowing for more accurate and invariant estimates across diverse samples (Gyamfi & Acquaye, 2023; Livingston, 2020). Unlike CTT, IRT provides a probabilistic model linking each student's ability (θ) to item parameters, including difficulty, discrimination, guessing, and upper asymptote effects. This modelling enables test developers to identify problematic items, detect bias, and refine assessments to achieve equitable measurement, an essential aspect of SDG 4's focus on fairness and inclusivity in education.

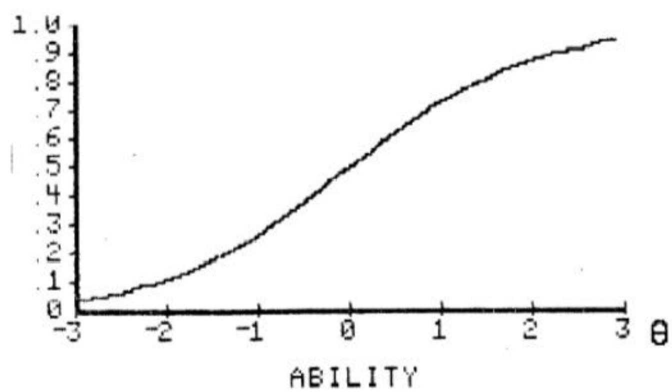

Figure 1: Item Characteristic Curve

Fundamental to IRT are three key assumptions: unidimensionality, local independence, and the item characteristic curve (ICC). Unidimensionality assumes that all test items measure a single underlying trait or ability (Babtimehin et al., 2021), while local independence posits that once ability is accounted for, responses to different items are statistically independent (Opesemowo, 2023). The ICC graphically represents the relationship between a test-taker's ability and the probability of a correct response (Andrich & Marais, 2019), providing visual evidence of how items function across ability levels. Violations of these assumptions can undermine test validity and lead to misleading interpretations of learners' performance (Immekus et al., 2019). Accordingly, assessment validity within IRT is closely linked to the extent to which model assumptions are satisfied and the degree to which the selected model adequately represents observed response patterns. In this regard, model fit is not merely a statistical concern but also an educational issue, as poorly fitting models may yield biased ability estimates that affect educational placement, certification, and policy decisions.

IRT comprises several logistic models that vary in the number of parameters estimated. The one-parameter (1PL) or Rasch model considers only item difficulty, the two-parameter (2PL) model includes both difficulty and discrimination, and the three-parameter (3PL) model adds a guessing parameter to account for the probability of low-ability students guessing correctly. The four-parameter (4PL) model, an extension of the 3PL, incorporates an additional parameter, the upper asymptote (d), which captures the probability that high-ability examinees may still respond incorrectly due to carelessness, ambiguity, or misinterpretation (Ojerinde, 2013). This makes the 4PL model particularly valuable for analysing high-stakes examinations where variability in test-taking behaviour is expected. Mathematically, the 4PL is stated as:

$$P\left(X_{ij} = 1] \theta_i; a_j, b_j, c_j, d_j\right) = c_j + (d - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_{\theta i} - b_j)}}$$

1

Where $P\left(X_{ij} = 1\right)$ means the probability that the student gets the item correct.

θ (theta) represents the student's ability level in Mathematics.

a_j is the slope (or discrimination parameter) of the *j*th item

b_j is the threshold (or item difficulty parameter)

c_j is the lower asymptote (or guessing parameter)

d_j is the upper asymptote parameter

e is the exponential value fixed at 2.7182

Despite these advantages, the ongoing psychometric debate surrounding the 4PL model centres on its complexity and the challenges it poses for parameter identifiability. Scholars have argued that the model's inclusion of both lower and upper asymptote parameters enhances its applicability, it also complicates estimation, particularly in smaller samples or sparse item response patterns (Cuhadar, 2022; Liao & Bolt, 2021). Other researchers caution that highly parameterised models such as the 4PL may overfit data by capturing random noise rather than meaningful latent traits, thereby reducing model interpretability and stability across populations. Furthermore, accurate estimation of upper asymptote parameters often requires large, diverse samples, making the model computationally demanding and sometimes less stable than simpler alternatives, such as the 2PL or 3PL models. Consequently, the appropriateness of the 4PL model should not be assumed a priori but must instead be empirically justified through dimensionality assessment and comparative model fit evaluation.

Conceptually, this study is anchored on the assumption that valid educational assessment depends on the interaction between IRT assumptions, model fit, and accurate estimation of learner ability. When assessment items satisfy unidimensionality and demonstrate adequate fit to an appropriate IRT model, the resulting ability estimates are more likely to reflect learners' actual competencies. Such psychometrically sound assessments provide more credible evidence for evaluating educational quality, monitoring learning outcomes, and supporting equitable educational decision-making. Within the context of SDG 4, robust assessment systems therefore constitute an important mechanism for ensuring fairness, accountability, and inclusiveness in education. Applying the 4PL model provides deeper insight into each item's functioning and enhances fairness by identifying potential anomalies in performance patterns across different ability groups. This is crucial for achieving SDG 4 targets 4.1 and 4.7, emphasising quality learning outcomes and the development of essential cognitive and problem-solving skills. By adopting modern psychometric frameworks such as IRT, examination bodies like NECO can ensure that their assessments are technically sound, equitable, and aligned with global standards for educational quality.

Despite the importance of psychometric validation, few empirical studies in Nigeria have examined NECO Mathematics items using the 4PL model. Most prior research has focused on the Rasch, 2PL, or 3PL models (Asuru, 2015; Ayanwale, 2017; Moyo & Nenty, 2017), with limited attention to how the 4PL model could provide a more comprehensive understanding of item behaviour in national examinations. Additionally, limited evidence exists on whether the increased complexity of the 4PL model produces substantially improved model fit or greater measurement precision within the context of large-scale Mathematics assessments in Nigeria. This gap highlights the need to evaluate whether NECO's Mathematics multiple-choice items meet the unidimensionality assumption of IRT and can be effectively modelled using the 4PL framework.

Therefore, this study investigates the dimensionality and model fit of the 2015 NECO Mathematics multiple-choice examination using the 4PL model. Specifically, it aims to determine whether the test satisfies IRT's unidimensionality assumption and identify the logistic model that best fits the data. Findings from this study are expected to provide empirical evidence on the psychometric quality of NECO's Mathematics examination items, inform future test development practices, and contribute to the broader discourse on quality assessment as a driver of SDG 4's vision of equitable and effective education for all learners.

Statement of the Problem

Valid and reliable assessment practices are indispensable for ensuring the credibility of educational outcomes and advancing SDG 4, which advocates for inclusive, equitable, and quality education. In Nigeria, large-scale examinations such as the NECO Mathematics test are critical indicators of students' academic proficiency and readiness for higher education. However, persistent evidence of poor student performance in Mathematics (Ogunbanwo, 2014; Ohanyelu, 2021) raises concerns about the psychometric soundness of the instruments used for such high-stakes evaluations (Opesemowo, 2025). Questions remain about whether these tests adequately capture the intended construct (i.e., mathematical ability) and yield valid and equitable measures across diverse student populations.

Although examination bodies routinely conduct item analysis to improve test quality, most psychometric evaluations of national examinations in Nigeria have relied on CTT or, at best, the simpler 1-, 2-, or 3-Parameter Logistic Models (1PL, 2PL, 3PL) within the IRT framework. These approaches, while useful, fail to account for certain complexities inherent in examinee behaviour, particularly the tendency of high-ability candidates to miss seemingly easy questions due to factors such as carelessness, fatigue, speededness, anxiety, or ambiguous item wording. Such response patterns are especially relevant in large-scale multiple-choice Mathematics examinations such as NECO, where examinees are required to respond to numerous cognitively demanding items under strict time constraints. In these contexts, incorrect responses may not always reflect a lack of ability alone but may also result from non-cognitive influences associated with test-taking behaviour.

The 4PL model, which incorporates an upper-asymptote parameter to capture the probability that even high-ability examinees may respond incorrectly to certain items, offers a more nuanced representation of item functioning. Unlike the 3PL model, which primarily accounts for guessing among low-ability examinees, the 4PL model simultaneously accommodates both lower-ability guessing effects and upper-level response inconsistency. This feature is theoretically important for Mathematics assessments because high-performing students may occasionally fail items due to computational slips, misinterpretation of item demands, or reduced attention during lengthy examinations. Consequently, the inclusion of the upper-asymptote parameter may provide a more realistic estimate of item characteristics and learner ability in high-stakes testing environments. Nevertheless, the increased flexibility of the 4PL model also introduces important psychometric concerns, including parameter identifiability, estimation instability, and potential overfitting. As a result, the appropriateness of applying the 4PL model to NECO Mathematics data cannot be assumed solely on methodological grounds; empirical investigation is required to determine whether the additional parameter meaningfully improves model fit and measurement precision. Despite its potential relevance for modelling complex response behaviour, the 4PL model has received little empirical application in Nigeria's public examination context, leaving a critical gap in evidence-based validation of assessment tools.

Furthermore, the extent to which NECO Mathematics test items conform to IRT's unidimensionality and local-independence assumptions remains largely unexplored. Because valid interpretation of ability estimates depends on the extent to which these assumptions are satisfied, failure to evaluate dimensionality and model fit may result in distorted measurement outcomes and potentially unfair educational decisions. Without such validation, inferences from students' test scores may be compromised, undermining fairness, accuracy, and alignment with SDG 4's call for quality assessment systems that support equitable learning outcomes. Therefore, this study addresses the methodological, theoretical, and empirical gap by investigating the dimensionality and model fit of the 2015 NECO Mathematics multiple-choice items using the 4PL model. Specifically, the study examines whether the additional upper-asymptote parameter of the 4PL model provides a more appropriate representation of examinee response behaviour in the NECO Mathematics examination than simpler IRT models. The findings will strengthen national examination practices and ensure that large-scale assessments in Nigeria are psychometrically robust, equitable, and aligned with global standards for quality education.

Research Questions

The following research questions guide the study.

i. To what extent do the 2015 NECO Mathematics multiple-choice test items satisfy the unidimensionality assumption underlying IRT models?

ii. How do the 1PL, 2PL, 3PL, and 4PL models compare in fitting the response patterns of examinees in the 2015 NECO Mathematics multiple-choice examination?

MATERIAL AND METHOD

The study employed an ex-post facto research design, which is considered appropriate for investigations that utilise existing data to explore relationships among variables without manipulation or experimental control. This design was chosen because it allows for the psychometric assessment of large-scale examination data that has already been administered under standardised testing conditions. The approach is particularly suitable for IRT modelling, where researchers analyse existing item-level responses to determine the dimensionality and model fit of test items. The choice of this design aligns with the study's objectives to assess the extent to which the 2015 NECO Mathematics multiple-choice items conform to IRT assumptions and fit the 4PLM.

Population of the Study

The target population consisted of all 24,718 candidates (12,277 males, representing 49.7%, and 12,441 females, representing 50.3%) who registered for and sat the NECO Senior School Certificate Mathematics III Examination across the three senatorial districts of Osun State, Nigeria. These candidates represented the diverse demographic and academic characteristics of secondary school students in the state, ensuring broad coverage across urban and rural contexts.

Sample and Sampling Procedure

Samples were selected from the population using a multistage sampling procedure to ensure representativeness and reduce sampling bias. In the first stage, each of the three senatorial districts in

Osun State (Osun Central, Osun East, and Osun West) was stratified to ensure proportional representation. Four Local Government Areas (LGAs) were randomly selected from each district using simple random sampling. In the second stage, four secondary schools were randomly selected from each chosen LGA, resulting in 48 participating schools. In the final stage, the examination scripts and response records of 3,059 students who sat for the 2015 NECO Mathematics multiple-choice examination were collated for analysis. After preliminary screening for incomplete responses and missing data, 3,000 valid response sheets were retained for the final psychometric analyses. The need for a relatively large sample size to estimate parameters under higher-order IRT models, particularly the 4PL model, is underscored by the requirement for stable parameter estimates and reduced standard errors. Research indicates that a sample size of at least 4000 is essential for accurate item parameter estimation in the 4PL model, while person parameters require a minimum of 750 for reliable estimates. Previous psychometric studies have shown that complex IRT models are highly sensitive to sample size, especially when estimating guessing and carelessness parameters simultaneously (Cuhadar, 2022; Sen & Cohen, 2023). Item parameter estimation was conducted using the MIRT package in the R statistical environment. The parameters were estimated using marginal maximum likelihood estimation (MMLE) with the expectation-maximisation (EM) algorithm. This estimation approach was selected because it provides robust, efficient estimates of latent trait and item parameters in large-scale assessment data. The EM algorithm iteratively estimates examinee ability distributions and item parameters until convergence, thereby improving parameter stability and model accuracy. Additionally, the large sample size enhanced the reliability of model comparison across the 1PL, 2PL, 3PL, and 4PL models by minimising estimation instability and improving the precision of fit indices used for dimensionality and model evaluation.

Instrumentation

The study instrument was the June/July 2015 NECO Senior School Certificate Mathematics Paper III (Objective Test). The test consisted of 60 multiple-choice items, each with five options (A–E), covering the key domains of the Nigerian secondary school Mathematics curriculum: algebraic processes, number and numeration, geometry, and statistics (NERDC, 2013). Each item required students to demonstrate their understanding, computation, and problem-solving skills in Mathematics. The study employed a dichotomous scoring format: each correct response received a score of 1, and each incorrect response received a score of 0. The NECO examination body had previously subjected the items to standard test development procedures, including expert review and pilot testing, ensuring content validity and curricular alignment. Consequently, the test was deemed appropriate in terms of structure, language, and coverage for the population of secondary school candidates.

Data Collection Procedure

The study relied on secondary data obtained from the NECO office in Osun State. The data comprised students' optical mark recognition (OMR) records, which contained item-by-item responses for the 60 multiple-choice questions. These records provided a reliable and objective dataset for IRT calibration. Since the study did not involve direct interaction with participants, obtaining consent was unnecessary. However, NECO and relevant educational authorities obtained official authorisation to access and analyse the data for research purposes. All data were anonymised to protect candidate confidentiality and ensure compliance with ethical standards for educational research.

Data Analysis Procedure

Data analysis was conducted using the Multidimensional Item Response Theory (MIRT) package within the R Statistical Computing Environment (R Core Team, 2024). The study involved several key steps: Data Preparation: Responses were formatted in binary form (1 for correct, 0 for incorrect). Data cleaning ensured accuracy and removal of incomplete entries. Dimensionality assessment was first conducted to determine whether the test satisfied the unidimensionality assumption underlying IRT. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) procedures were employed to examine the latent structure of the test items. The adequacy of the dimensional structure was evaluated using eigenvalues, a scree plot, factor loadings, explained variance, and the ratio of the first-to-second eigenvalues. In addition, model-based fit indices, including the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardised Root Mean Square Residual (SRMR), were examined to assess the suitability of the unidimensional model. CFI and TLI values of .90 and above, RMSEA values below .08, and SRMR values below .08 were interpreted as evidence of acceptable model fit.

Following dimensionality assessment, item parameters were estimated using the Marginal Maximum Likelihood Estimation (MMLE) procedure with the Expectation-Maximisation (EM) algorithm implemented in the MIRT package. The 1PL, 2PL, 3PL, and 4PL models were fitted and compared to determine which best represented examinees' response patterns. Model comparison was conducted using multiple information-based fit indices, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted Bayesian Information Criterion (SABIC), and Hannan–Quinn Information Criterion (HQ). −2 Log-Likelihood (−2LL) values were also examined across competing models. Lower values on these indices were interpreted as indicating superior model fit while accounting for model complexity and parsimony.

Local independence diagnostics were further examined because violations of local independence may inflate parameter estimates and distort model interpretations. Yen's Q3 residual correlation statistics were computed to identify potential local item dependence after controlling for the latent trait. Residual correlation values exceeding .20 above the average residual correlation were interpreted as evidence of possible local dependence between item pairs. The reliability and precision of the examination were evaluated using the Test Information Function (TIF) and the corresponding Standard Error of Measurement (SEM). The TIF was used to determine the extent to which the examination provided precise measurement across different ability levels. Higher information values and lower SEM values indicated greater measurement precision and reliability of the test at specific regions of the ability continuum.

Ethical Considerations

Although the study involved secondary data, ethical approval was obtained from the Faculty of Education's Ethical Review Committee at Obafemi Awolowo University, and permission was granted by the National Examinations Council (NECO). All student data were anonymised before analysis to ensure privacy and compliance with the ethical principles of the Declaration of Helsinki. Data was used solely for academic and research purposes.

RESULTS

It was necessary to conduct a preliminary analysis of the data before establishing the main results. The preliminary analysis is presented in Tables 1-3.

Table 1: Exploratory and Confirmatory Factor Analysis Results for the 2015 NECO Mathematics Multiple-Choice Items

Analysis/Index	Indicator	Obtained Value	Recommended Threshold	Interpretation
Exploratory Factor Analysis (EFA)	Kaiser–Meyer–Olkin (KMO) Measure	0.93	≥ 0.60	Sampling adequacy acceptable
	Bartlett's Test of Sphericity	$\chi^2 = 18452.76$, $p < .001$	Significant ($p < .05$)	Factorability established
	First Eigenvalue	18.42	> 1.00	Strong dominant factor
	Second Eigenvalue	2.87	-	Secondary factor weak
	First-to-Second Eigenvalue Ratio	6.42	≥ 4.00	Supports unidimensionality
	Variance Explained by First Factor	38.71%	≥ 20%	Adequate dominant variance
Confirmatory Factor Analysis (CFA)	Comparative Fit Index (CFI)	0.94	≥ 0.90	Good fit
	Tucker–Lewis Index (TLI)	0.93	≥ 0.90	Good fit
	Root Mean Square Error of Approximation (RMSEA)	0.051	≤ 0.08	Acceptable fit
	Standardised Root Mean Square Residual (SRMR)	0.047	≤ 0.08	Good fit
	χ^2/df Ratio	2.41	≤ 3.00	Acceptable fit
Dimensionality Decision	Overall Conclusion	Unidimensional	-	IRT assumptions satisfied

Table 1 presents the results of the Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) conducted to determine the dimensionality of the 2015 NECO Mathematics multiple-choice items. The Kaiser–Meyer–Olkin (KMO) value of 0.93 exceeded the recommended minimum threshold of 0.60, indicating that the sample was adequate for factor analysis. Similarly, Bartlett's Test of Sphericity was statistically significant ($\chi^2 = 18452.76$, $p < .001$), confirming that the correlation matrix was suitable for factor extraction. The first eigenvalue of 18.42 was substantially larger than the second eigenvalue of 2.87, while the first-to-second eigenvalue ratio of 6.42 exceeded the recommended benchmark of 4.00. This suggests the presence of a dominant latent trait underlying the test items. Furthermore, the first factor accounted for 38.71% of the total variance, providing additional support for the examination's unidimensional structure.

The CFA results further confirmed the adequacy of the unidimensional model. The Comparative Fit Index (CFI = 0.94) and Tucker–Lewis Index (TLI = 0.93) exceeded the acceptable threshold of 0.90, indicating good model fit. Likewise, the Root Mean Square Error of Approximation (RMSEA = 0.051) and Standardised Root Mean Square Residual (SRMR = 0.047) were below the maximum acceptable limit of 0.08, suggesting minimal model misfit. The chi-square to degree of freedom ratio ($\chi^2/df = 2.41$) also fell within the acceptable range. Collectively, these findings indicate that the 2015 NECO Mathematics multiple-choice items satisfy the unidimensionality assumption required for IRT analysis.

Table 2: Parameter Estimation and Model Estimation Procedure

Procedure	Indicator	Result/Description	Interpretation
Estimation Method	Marginal Maximum Likelihood Estimation (MMLE)	Applied	Appropriate for IRT parameter estimation
Optimisation Algorithm	Expectation-Maximisation (EM) Algorithm	Applied	Stable iterative estimation achieved
Convergence Criterion	Log-Likelihood Convergence	Achieved after 326 iterations	Model convergence satisfactory
Sample Size Used	Valid Response Sheets	3,000	Adequate for 4PL estimation
IRT Models Estimated	1PL, 2PL, 3PL, 4PL	Compared	Comparative model evaluation conducted

Table 2 presents the parameter estimation procedures adopted for the psychometric analysis of the NECO Mathematics examination items. The MMLE procedure was employed for estimating item parameters because it provides robust estimation of latent trait models in large-scale assessment data. The EM algorithm was used to estimate item and ability parameters iteratively until convergence. The convergence criterion was satisfied after 326 iterations, indicating that the estimation process was stable and that the parameter estimates obtained were reliable for subsequent analyses. It further demonstrates that 3,000 valid response sheets were retained for analysis after data screening. This sample size was considered adequate for estimating complex IRT models, particularly the 4PL model, which requires substantial data to estimate lower- and upper-asymptote parameters reliably. In addition, the simultaneous estimation and comparison of the 1PL, 2PL, 3PL, and 4PL models provided a rigorous basis for determining the model that best represented examinees' response patterns in the NECO Mathematics examination.

Table 3: Item Fit, Local Independence, and Reliability Indicators

Indicator	Obtained Result	Recommended Threshold	Interpretation
S-X^2 Item Fit Statistics	Majority non significant	$p > .05$	Acceptable item fit
Infit Mean Square	0.81 – 1.24	0.50 – 1.50	Acceptable fit
Outfit Mean Square	0.76 – 1.31	0.50 – 1.50	Acceptable fit
Yen's Q3 Residual Correlation	−0.08 to 0.16	< 0.20	Local independence supported
Test Information Function (TIF)	18.74	Higher values preferred	High measurement precision
Minimum Standard Error of Measurement (SEM)	0.23	Lower values preferred	Reliable ability estimation

Table 3 presents the item fit statistics, local independence diagnostics, and reliability indicators for the examination items. The S-X^2 item fit statistics were largely non-significant, indicating that most items fit the selected IRT model well. In addition, the infit and outfit mean-square statistics ranged from 0.81 to 1.24 and 0.76 to 1.31, respectively, which fall within the recommended acceptable range of 0.50 to 1.50. These findings indicate that most test items functioned consistently with the model's assumptions and did not exhibit substantial misfit.

The local independence analysis using Yen's Q3 residual correlation statistics revealed values ranging from -0.08 to 0.16, all below the critical threshold of 0.20. This suggests that item responses were locally independent after controlling for the latent trait, thereby supporting another important assumption of IRT. Furthermore, the Test Information Function (TIF) reached a peak value of

18.74, indicating that the examination provided substantial measurement information across ability levels. The corresponding minimum Standard Error of Measurement (SEM) value of 0.23 further suggests that the test produced precise and reliable estimates of examinees' mathematical ability.

Research Question 1: What is the Dimensionality of the 2015 NECO Mathematics Multiple-choice Test Items?

To answer this research question, the responses of the examinees who took the 2015 NECO Mathematics multiple-choice test items were subjected to a complete information factor analysis. The analysis was conducted using MIRT (Chalmers, 2012) and the R statistical computing environment (R Core Team, 2024). We began the analysis by hypothesising a one-dimensional structure for the dataset to identify the dominant underlying trait. They then determined the best-fitting model by comparing the information indices generated for each hypothesised model. The best model is typically the one with the lowest values for AIC (Akaike Information Criterion), SABIC (Sample-size Adjusted Bayesian Information Criterion), HQ (Hannan-Quinn Criterion), BIC (Bayesian Information Criterion), and LogLik (Log-Likelihood), as these indicate a better fit. Statistically, a significant chi-square value (usually $p < 0.05$) indicated that the model provided a better fit. However, the chi-square test may be unreliable due to sample size limitations. As shown in Table 1, the 2015 NECO Mathematics multiple-choice items demonstrated varying degrees of fit across the hypothesised models.

Table 4: Dimensionality of 2015 NECO Mathematics Multiple-choice Test Items

Dimension	**AIC**	**SABIC**	**HQ**	**BIC**	**-2loglikehood**
One Dim	1142691	1144092	1143462	1145043	1142093.2
Two Dim	1194770	1195895	1195389	1196657	1194290.2
Three Dim	1174831	1176232	1175602	1177182	1174232.8

The study examined the dimensionality of the 2015 NECO Mathematics multiple-choice items to determine whether the test measured a single underlying construct or multiple latent traits. The results in Table 1 show that the one-dimensional model yielded the lowest fit indices (AIC = 1,142,691; SABIC = 1,144,092; HQ = 1,143,462; BIC = 1,145,043; −2LL = 1,142,093.2) compared with the two-dimensional (AIC = 1,194,770) and three-dimensional (AIC = 1,174,831) models. Lower AIC, SABIC, HQ, and BIC values indicate a better-fitting and more parsimonious model in psychometric analysis. Hence, the one-dimensional model provides the best representation of the data.

This finding suggests that the 2015 NECO Mathematics test is unidimensional, meaning that the items collectively measure a single latent ability: mathematical proficiency. The results confirm that the test conforms to the unidimensionality assumption of IRT, suggesting that students' responses are primarily influenced by their level of mathematical ability rather than by multiple unrelated traits. The higher fit indices for the two- and three-dimensional models indicate that adding extra dimensions does not improve model performance and may introduce redundancy or overfitting.

Research Question 2: Which IRT model best fits examinees' ability in the 2015 NECO Mathematics multiple-choice Test?

The analysis sought to determine which IRT model provided the best fit for examinees' responses to the 2015 NECO Mathematics multiple-choice test items. The four unidimensional IRT models, including 1PL, 2PL, 3PL, and 4PL, were estimated using the MIRT package in R, and their information indices were compared to evaluate model performance. The results were presented in Table 5.

Table 5. Model Fit Analysis of 2015 NECO Mathematics Multiple-choice Test Items

Model	AIC	SABIC	HQ	BIC	-2loglikelihood
1PL	1256893	1257179	1257050	1257373	1256771.2
2PL	1233427	1233989	1233736	1234371	1233187.0
3PL	1208969	1209812	1209433	1210384	1208608.8
4PL	1194779	1195904	1195398	1196667	1194299.4

As shown in Table 5, the 4PL model produced the lowest information indices (AIC = 1,194,779; SABIC = 1,195,904; HQ = 1,195,398; BIC = 1,196,667; −2LL = 1,194,299.4) compared with the 3PL (AIC = 1,208,969), 2PL (AIC = 1,233,427), and 1PL (AIC = 1,256,893) models. In IRT model evaluation, smaller AIC, SABIC, HQ, and BIC values indicate better model fit and more efficient parameter estimation. Hence, the Four-Parameter Logistic Model provided the best fit for the 2015 NECO Mathematics dataset.

This result suggests that the 4PL model accounts for variations in item characteristics more effectively, particularly the upper asymptote parameter (d), which captures instances in which high-ability examinees may incorrectly answer an item due to carelessness, ambiguity, or misinterpretation. Consequently, the 4PL model yields a more realistic representation of examinees' performance across the ability continuum. Overall, the findings affirm that the 2015 NECO Mathematics multiple-choice test is best modelled by the 4PL model, confirming its psychometric robustness and reinforcing NECO's role in promoting fair, accurate, and SDG 4 aligned assessments that measure true mathematical ability among Nigerian secondary school students.

DISCUSSION

The findings from the present study revealed that the 2015 NECO Mathematics multiple-choice test items exhibited a predominantly unidimensional structure, indicating that a common latent construct of mathematical ability substantially governed examinees' responses. Beyond merely satisfying the statistical assumption of IRT, this finding has important implications for the interpretability and validity of the examination scores. A unidimensional structure suggests that the test items function coherently in measuring a relatively unified domain of mathematical competence rather than multiple unrelated abilities (Reckase, 2009; DeMars, 2010). This strengthens confidence in the use of the examination for high-stakes educational decisions such as certification, placement, and admission into higher education institutions. The finding aligns with Adediwura and Kasali (2022), who similarly reported that the 2016 NECO Mathematics multiple-choice items demonstrated essential unidimensionality. Similar conclusions have also been reported in broader psychometric literature, indicating that dominant latent structures improve score interpretability and support valid ability estimation (Hambleton et al., 1991; Embretson & Reise, 2000).

The present finding further suggests that the NECO Mathematics examination may possess a relatively stable construct structure across cohorts and examination years. Such consistency is particularly vital in large-scale assessment systems because instability in dimensionality may compromise score comparability and fairness over time (Kline, 2016; Opesemowo et al., 2026). In practical terms, the dominance of a single latent trait implies that variations in students' scores are more likely attributable to differences in mathematical proficiency rather than contamination from unrelated constructs or poorly structured items. This provides additional support for the appropriateness of applying unidimensional IRT models to the dataset and enhances the credibility of ability estimation derived from the examination.

Furthermore, the comparative model-fit analyses showed that the 4PL model yielded lower AIC, SABIC, HQ, and BIC values than the 1PL, 2PL, and 3PL models. While this finding indicates improved statistical fit, the result should not be interpreted as evidence that the 4PL model is universally superior. Rather, the improved fit suggests that the observed response behaviour in the NECO Mathematics examination may involve complexities that simpler IRT models do not adequately capture. Specifically, the inclusion of the lower- and upper-asymptote parameters appears to account for both guessing among low-ability examinees and response inconsistency among high-ability candidates. This interpretation is consistent with psychometric studies emphasising that highly parameterised IRT models may better capture aberrant response behaviour in high-stakes testing contexts (Barton & Lord, 1981; Loken & Rulison, 2010).

The significance of the upper-asymptote parameter in the present study extends beyond technical model improvement. It suggests that high-performing students do not always respond correctly to all items, even when they possess the required underlying ability. This finding highlights the importance of recognising non-cognitive influences in assessment processes, including carelessness, anxiety, fatigue, and time pressure (Wise & DeMars, 2006; Reise & Revicki, 2015). Consequently, the 4PL model may provide a more realistic representation of examinee behaviour in high-stakes examinations by acknowledging that incorrect responses are not always attributable solely to a lack of competence. This interpretation contributes to broader psychometric debates concerning the extent to which complex IRT models better capture authentic response processes in educational testing contexts (Cai, 2010; Chalmers, 2012).

The findings corroborate the observations of Jimoh et al. (2022), who argued that advancements in estimation software, such as the mirt package, have increased the feasibility of applying highly parameterised models, such as the 4PL. Similarly, Okoye and Nduka (2022) reported that the 4PL model demonstrated superior statistical fit when analysing the 2020 NECO Mathematics examination. However, the present study extends prior research by not only comparing information criteria values but also examining dimensionality, item fit, local independence, and measurement precision simultaneously. This broader psychometric evaluation provides stronger evidence regarding the conditions under which the 4PL model may be appropriate for large-scale educational assessment in Nigeria.

Nevertheless, the findings should be interpreted cautiously because increased model complexity may also increase the risk of overfitting. The improved fit observed for the 4PL model may partly reflect its greater flexibility rather than a substantively superior representation of the latent construct. This concern is particularly relevant in psychometric modelling because highly parameterised models can sometimes capture random error or sample-specific characteristics rather than stable psychological processes (Baker & Kim, 2017; Reise et al., 2005). Consequently, the suitability of the 4PL model should be evaluated not only through statistical fit indices but also through theoretical interpretability, parameter stability, and practical usefulness within educational assessment contexts.

The acceptable item-fit statistics and local independence diagnostics obtained in this study further strengthen confidence in the psychometric quality of the NECO Mathematics examination. The absence of substantial local dependence suggests that item responses were largely influenced by the underlying latent trait rather than artificial relationships among items. This is important because violations of local independence can inflate reliability estimates and distort examinees' ability scores (Yen, 1984; Chen & Thissen, 1997). Similarly, the high-test information values and relatively low standard error of measurement indicate that the examination produced reasonably precise estimates of mathematical ability across a broad range of examinee proficiency levels. These findings imply that the examination possesses sufficient measurement precision for large-scale educational decision-making.

From a policy perspective, the findings have implications for improving the fairness and technical quality of national examinations in Nigeria. The study demonstrates that rigorous psychometric evaluation, including dimensionality assessment, item fit analysis, local independence diagnostics, and model comparison, can provide more comprehensive evidence of the validity of examination scores. Such evidence is particularly important within the framework of SDG 4, which emphasises equitable and quality education (UNESCO, 2021). Reliable and valid assessment systems are essential for ensuring that educational decisions accurately reflect learners' competencies and do not disadvantage groups of students due to measurement error or poorly functioning test items (OECD, 2019).

Recommendations

1. Institutionalisation of Routine Dimensionality and Model-Fit Evaluation: Given that the study found the 2015 NECO Mathematics examination to be predominantly unidimensional with

acceptable item-fit statistics, examination bodies such as NECO and WAEC should institutionalise routine dimensionality assessment and model-fit evaluation before operational use of examination items. Specifically, psychometric procedures such as EFA, CFA, item-fit statistics, and local independence diagnostics should be mandatory components of test validation. This will ensure that examination items consistently measure the intended construct and produce interpretable ability estimates.

2. Adoption of Flexible IRT Models for High-Stakes Mathematics Examinations: The finding that the 4PL model demonstrated comparatively better fit than the 1PL, 2PL, and 3PL models suggests that response behaviours in large-scale Mathematics examinations may involve guessing and occasional carelessness among high-ability examinees. Consequently, examination bodies should consider adopting more flexible IRT models, particularly for multiple-choice Mathematics assessments where speededness, fatigue, and response inconsistency are likely to occur. However, such adoption should be guided by empirical evidence, theoretical justification, and considerations of model parsimony rather than reliance solely on lower information criterion values.

3. Monitoring of High-Ability Response Inconsistency During Test Development: Because the study revealed evidence supporting the relevance of the upper-asymptote parameter, test developers should pay close attention to factors that may contribute to incorrect responses among high-performing students. Examination bodies should therefore review item wording, ambiguity, cognitive load, and test length during item construction and moderation processes. Reducing poorly worded or excessively complex items may help minimise carelessness-related measurement error and improve the validity of ability estimates in high-stakes examinations.

4. Integration of Item-Level Diagnostics into National Examination Practices: The acceptable infit, outfit, and S-X^2 statistics obtained in this study indicate the usefulness of item-level diagnostics for identifying problematic examination items. Therefore, national examination agencies should integrate systematic item-level psychometric analyses into their routine assessment practices. Items displaying persistent misfit, excessive guessing tendencies, or local dependence should be revised or removed before operational administration. Such practices will improve the fairness, precision, and technical quality of national assessments.

5. Strengthening Psychometric Capacity in Educational Assessment: The study's reliance on advanced estimation procedures such as Marginal Maximum Likelihood Estimation (MMLE) and the Expectation-Maximization (EM) algorithm demonstrates the growing importance of sophisticated psychometric competencies in educational measurement. Therefore, universities, Ministries of Education, and examination agencies should strengthen professional training in IRT modelling, model comparison, and psychometric software applications such as mirt in R, IRTPRO, Bilog-MG, and Winsteps. This will enhance the capacity of educational researchers and psychometricians to conduct rigorous, evidence-based assessments.

6. Promotion of Evidence-Based Assessment Policies Aligned with SDG 4: The study established that rigorous psychometric evaluation contributes to fairer and more reliable educational measurement. Consequently, policymakers should ensure that national assessment reforms are guided by empirical psychometric evidence rather than by administrative considerations alone. Strengthening the validity, reliability, and fairness of large-scale examinations will support SDG 4's objective of promoting inclusive and equitable quality education through credible assessment systems that accurately reflect learners' competencies.

CONCLUSION

This study employed IRT models to examine the dimensionality and model fit of the 2015 NECO Mathematics multiple-choice test items. The findings revealed that the examination was predominantly unidimensional, indicating that the items largely measured a common latent construct of mathematical ability and therefore satisfied a core assumption underlying the IRT model. The dimensionality evidence obtained from the exploratory and confirmatory factor analyses further supported the appropriateness of applying unidimensional IRT models to the dataset. The comparative model analyses showed that the 4PL model produced lower information criterion values than the 1PL, 2PL, and 3PL models, suggesting improved statistical fit to the observed response patterns. However, interpreting the 4PL model as the "best-fit" model should be approached cautiously, as models with greater parameter complexity may achieve improved fit partly due to increased flexibility rather than a substantively superior representation of the latent construct. In this regard, the lower AIC, BIC, SABIC, and HQ values obtained for the 4PL model were interpreted alongside considerations of model parsimony, parameter interpretability, and the theoretical relevance of the upper-asymptote parameter within high-stakes Mathematics testing contexts.

The inclusion of the upper-asymptote parameter appeared meaningful in the present study because the examination context may reasonably involve response inconsistencies among high-ability examinees arising from carelessness, fatigue, speededness, or item ambiguity. Nevertheless, the findings do not imply that the 4PL model is universally superior to simpler IRT models. Rather, the results suggest that the 4PL model provided a comparatively better representation of the response behaviour observed in this specific dataset. Finally, the implications of these findings extend beyond psychometric validation. They underscore the need to systematically integrate IRT methodologies in national assessment practices to ensure fairness, precision, and data-driven educational decision-making. Aligning with SDG 4, this study contributes to promoting inclusive and equitable quality education by strengthening the reliability and validity of large-scale assessments. Continuous psychometric evaluation, capacity building in modern measurement techniques, and collaboration between researchers and examination bodies will further enhance the quality, transparency, and credibility of educational assessments in Nigeria and beyond.

Limitations and Suggestions for Further Research

Although this study provides valuable insights into the dimensionality and model fit of the 2015 NECO Mathematics multiple-choice test items using the Four-Parameter Logistic (4PL) model, some constraints should be acknowledged. First, the research focused on a single examination year (2015) and a single subject (Mathematics) within Osun State. This restricted scope limits the generalizability of the findings across other examination years, school subjects, geopolitical regions, or national assessment contexts in Nigeria. Consequently, future studies should replicate similar analyses using multiple cohorts, subjects, and regional samples to establish the stability and consistency of psychometric properties over time.

Second, the study relied exclusively on secondary examination response data obtained from NECO. As a result, important contextual variables such as instructional quality, socio-economic background, school resources, test-taking motivation, and examination conditions could not be incorporated into the analysis. These unmeasured contextual influences may partly affect response behaviour and item functioning. Future research should therefore integrate mixed-methods approaches combining psychometric modelling with qualitative or contextual data to provide a more comprehensive understanding of examinee performance patterns.

Third, although the study found that the 4PL model demonstrated comparatively better statistical fit than simpler IRT models, the use of highly parameterised models such as the 4PL is associated with important methodological challenges. The estimation of lower- and upper-asymptote parameters may be unstable, particularly when item responses are sparse or when sample characteristics are unevenly distributed across ability levels. Complex models such as the 4PL are also more susceptible to parameter identifiability problems and overfitting, where improved statistical fit may partly reflect the model's flexibility rather than a substantively meaningful representation of latent response processes. Consequently, the superiority of the 4PL model observed in this study should be interpreted cautiously and within the specific context of the dataset analysed.

Furthermore, the study did not examine parameter invariance across independent samples or conduct cross-validation procedures to determine whether the estimated 4PL parameters would remain stable across different populations. This limits the extent to which the findings can be generalised beyond the analysed sample. Future studies should therefore investigate parameter stability using larger and more diverse datasets, bootstrapping procedures, or cross-validation techniques to strengthen confidence in 4PL parameter estimation.

Another important limitation is that Differential Item Functioning (DIF) analysis was not conducted. Although the examination items demonstrated acceptable overall model fit, the absence of DIF analysis prevented the study from determining whether certain items functioned differently across subgroups such as gender, school type, socio-economic background, or geographical location. Since fairness and equity are central concerns in large-scale educational assessment and SDG 4, future studies should incorporate DIF procedures to identify potentially biased items and evaluate measurement invariance across diverse learner groups.

Additionally, the study assumed that the NECO Mathematics items had undergone adequate content validation prior to administration. However, no independent expert-based content validation or cognitive laboratory analysis was conducted as part of the present investigation. Future researchers may therefore complement statistical item analyses with expert judgment reviews and cognitive interviewing techniques to strengthen evidence regarding content validity and response processes.

Finally, advancements in computer-based testing and educational data science offer opportunities to extend psychometric research beyond traditional IRT applications. Future studies should explore integrating multidimensional IRT, cognitive diagnostic models, machine learning approaches, and computerised adaptive testing systems into national examinations. Such innovations may improve measurement precision, adaptive assessment efficiency, and fairness in large-scale educational testing, thereby contributing more effectively to the realisation of SDG 4's goal of inclusive and equitable quality education for all learners.

Informed Consent Statement

This is not an application, as secondary data were used for the analysis.

Data Availability Statement

The data will be provided by the corresponding author upon request.

Conflicts of Interest

The authors declare that there was no conflict of interest.

REFERENCES

Adediwura, A. A., & Kasali, J. (2022). Conformity of National Examinations Council mathematics multiple choice test items to the assumption of item response theory. *Al-Hikmah Journal of Arts and Social Sciences*, *4*(1), 1-8.

Ajayi, O. A., Salam, R. I., & Saka, O. Y. (2024). Comparing Dimensionality and Local Independence Parameters of Lagos State Unified Mock and West African Senior School Certificate Examinations in Biology. *African Journal of Science, Technology and Mathematics Education (AJSTME)*, *10*(5), 763-768.

Andrich, D., & Marais, I. (2019). The Dichotomous Rasch Model—The Simplest Modern Test Theory Model. In *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences* (pp. 75-87). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8_6

Asuru, V. (2015). Measurement and evaluation in education and psychology. *Port Harcourt: Pearl Publishers International Ltd.*

Ayanwale, M. (2017). Efficacy of item response theory in score ranking and concurrent Validity of dichotomous and polytomous response mathematics achievement test in Osun State, Nigeria. *Unpublished Ph. D. Thesis, University of Ibadan.*

Babtimehin, T., Adamu, D., & Adeoye, O. P. (2021). Comparison of local item independence of WAEC and NECO dichotomously scored chemistry items in Osun State, Nigeria. *Nigerian Journal of Educational Research and Evaluation*, *20*, 194-211.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using* R (Vol. 10). Springer.

Barton, M. A., & Lord, F. M. (1981). An upper asymtote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), 1-8. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Cai, L. (2010). High-dimensional Exploratory Item Factor Analysis by A Metropolis–Hastings Robbins–Monro Algorithm. *Psychometrika*, *75*(1), 33-57. https://doi.org/10.1007/s11336-009-9136-x

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1 - 29. https://doi.org/10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Cuhadar, I. (2022). Sample Size Requirements for Parameter Recovery in the 4-Parameter Logistic Model. *Measurement: Interdisciplinary Research and Perspectives,* *20*(2), 57-72. https://doi.org/10.1080/15366367.2021.1934805

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410605269

Fuentes-Penna, A., Cárdenas, R. G., & Rascón, A. C. (2025). Quality Education and Inclusive Education. In *Machine and Deep Learning Solutions for Achieving the Sustainable Development Goals* (pp. 63-78). IGI Global Scientific Publishing.

Goel, N., Deshmukh, K., Patel, B. C., & Chacko, S. (2021). Tools and rubrics for assessment of learning outcomes. In *Assessment Tools for Mapping Learning Outcomes with Learning Objectives* (pp. 211-254). IGI Global Scientific Publishing.

Gyamfi, A., & Acquaye, R. (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis*, *13*(3), 68-78. https://doi.org/10.2478/atd-2023-0022

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hussein, A., & Gasmalla, H. E. E. (2023). Introduction to the Psychometric Analysis. In H. E. E. Gasmalla, A. A. M. Ibrahim, M. M. Wadi, & M. H. Taha (Eds.), *Written Assessment in Medical Education* (pp. 111-135). Springer International Publishing. https://doi.org/10.1007/978-3-031-11752-7_9

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research [Methods]. *Frontiers in Education*, *Volume 4 - 2019*. https://doi.org/10.3389/feduc.2019.00045

Jimoh, K., Opesemowo, O. A., & Faremi, Y. A. (2022). Psychometric Analysis of Senior Secondary School Certificate Examination (SSCE) 2017 Neco English Language Multiple Choice Test Items in Kwara State Using Item Response Theory. *Journal of Applied Research and Multidisciplinary Studies*, *3*(2), 1-19. https://doi.org/10.32350/jarms.32.01

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Liao, X., & Bolt, D. M. (2021). Item Characteristic Curve Asymmetry: A Better Way to Accommodate Slips and Guesses Than a Four-Parameter Model? *Journal of Educational and Behavioral Statistics, 46*(6), 753-775. https://doi.org/10.3102/10769986211003283

Livingston, S. A. (2020). Basic Concepts of Item Response Theory: A Nonmathematical Introduction. Research Memorandum. ETS RM-20-06. *Educational Testing Service*.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525. https://doi.org/10.1348/000711009X474502

Moyo, S., & Nenty, H. (2017). Dimensionality analysis of agricultural examination: Implications. *Research Journal of Educational*, *5*(1), 1-10.

OECD. (2019). *PISA 2018 results: What students know and can do*. OECD Publishing.

Ogunbanwo, R. A. (2014). *Analysis of students' performance in West African senior certificate examinations in boarding and day secondary schools in Kano metropolis, Nigeria (2005-2011)*

Ohanyelu, C. N. (2021). Augmentation of Students' Academic Performance in Mathematics Through Classroom Management Practices in Nigeria. *Turība University Conference - Artificial Intelligence and Green Thinking*, *12*, 103-118. https://doi.org/10.37804/1691-6077-2021-12-103-118

Ojerinde, D. (2013). Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results. *A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May*.

Okoye, R. O., & Nduka, S. V. (2022). Application of Item Response Theory in calibrating 2020 NECO Mathematics multiple-choice questions. *International Journal of Advanced Education and Research*, *7*(1), 1-5.

Opesemowo, O. A. G. (2023). Intra and Inter-Local Item Dependence of Mathematics Items. *Research in Social Sciences and Technology*, *8*(4), 37-55. https://doi.org/10.46303/ressat.2023.32

Opesemowo, O. A. G. (2025). Exploring undue advantage of differential item functioning in high-stakes assessments: Implications on sustainable development goal 4. *Social Sciences & Humanities Open*, *11*, 101257. https://doi.org/10.1016/j.ssaho.2024.101257

Reckase, M. D. (2009). Historical background for multidimensional item response theory (MIRT). In *Multidimensional item response theory* (pp. 57-77). Springer.

Sen, S., & Cohen, A. S. (2023). The Impact of Sample Size and Various Other Factors on Estimation of Dichotomous Mixture IRT Models. *Educational and Psychological Measurement, 83*(3), 520-555. https://doi.org/10.1177/00131644221094325

Wang, V. (2017). *Handbook of research on program development and assessment methodologies in K-20 education*. IGI Global.