ITEM PARAMETERS, SCORING MODELS AND ABILITY ESTIMATES OF DISTANCE EDUCATION STUDENTS: IMPLICATIONS FOR THE NATIONAL OPEN UNIVERSITY OF NIGERIA

AMOS ILIYA, PhD
Regional Training and Research Institute for Distance and Open Learning
National Open University of Nigeria, Headquarters, Abuja
E-Mail: iamos@noun.edu.ng

Abstract

This study investigates item parameters, scoring models, and ability estimates of distance education students, with implications for the National Open University of Nigeria (NOUN). As the institution continues to expand its distance education offerings, understanding the effectiveness of assessment tools becomes increasingly crucial. The study employs both Classical Test Theory (CTT) and Item Response Theory (IRT) to analyze item parameters. Survey data were collected from a sample of 126 second-year Mathematics students at NOUN. Instruments used for the analysis include 1-Parameter Logistic Model (1PLM-MAT), 2-Parameter Logistic Model (2PLM-MAT), and 3-Parameter Logistic Model (3PLM-MAT). Data were analyzed using R and SPSS software packages to facilitate data interpretation. The findings reveal that polytomous scoring models provide more reliable estimates of student ability compared to dichotomous scoring models under the 2-PL and 3-PL models. The study recommends that NOUN consider integrating a variety of scoring models, including those incorporating partial credit scoring and polytomous item responses, to enable a more detailed assessment of student performance, particularly in disciplines where binary scoring may not fully reflect learning outcomes.

Keywords: Item Parameters, Scoring Models, Ability Estimates, Distance Education, Psychometric Analysis
DOI:https://doie.org/10.0913/SER.2024583552

Introduction

The assessment of students' performance in distance education has become increasingly significant within the context of Open and Distance Learning (ODL), particularly at institutions like the National Open University of Nigeria (NOUN). As the largest ODL institution in Nigeria, NOUN serves a diverse and widespread student population, making the accurate measurement of student abilities critical for maintaining academic standards and promoting effective learning. Since the purpose of testing is to estimate learners' abilities, i.e., latent traits or constructs, the use of item parameters, scoring models, and ability estimates is essential for creating reliable and valid assessments that accurately reflect the competencies of students across varied learning environments (Igbokwe & Ogili, 2020).

Item parameters—specifically difficulty, discrimination, validity, and reliability—are vital in the construction and evaluation of test items. These parameters help educators design assessments that are appropriately challenging, capable of distinguishing between students of different ability levels, and resistant to random guessing (Alabi, 2021). Scoring models, particularly those grounded in Classical Test Theory (CTT) and Item Response Theory (IRT) like the dichotomous and polytomous models, offer different methodologies for interpreting test data and estimating student abilities. While CTT provides a straightforward approach based on total scores, IRT offers a more sophisticated analysis by accounting for the interaction between test items and student abilities, thereby offering a deeper understanding of students' performance (Kpolovie, 2017).

IRT models examine performance at the item level, contrasting with CTT, which depends on test scores for ability estimation. IRT attempts to model the relationship between an unobserved variable, typically conceptualized as an examinee's ability, and the probability of the examinee correctly responding to a particular test item. Three widely used models are the 1-parameter logistic (Rasch Model), 2-parameter logistic, and 3-parameter logistic models. These models assume a single underlying ability (β) for examinees but vary in the characteristics they assign to test items (Tella & Bashorun, 2019). All three models include an item difficulty parameter (b), which represents the point of inflection on the ability (β) scale. For the 1-parameter and 2-parameter models, this is the point on the ability (β) scale where an examinee has a 50% chance of correctly answering an item (Adeoye, 2020).

According to Olawale and Fatokun (2022), the issue may lie less in item parameterization and more in the scoring models (right/wrong, partial credit, rating scale, etc.) that influence ability estimation. Early research on IRT focused primarily on dichotomous models, where test item responses are scored as either right or wrong (1, 0). Following Lord and Novick's establishment of the 1-, 2-, and 3-parameter logistic models for dichotomous items in 1953, Samejima introduced the first polytomous model, the Graded Response Model, in 1969. In 1972, Bock and Samejima presented the Nominal Categories Model, another polytomous model, but significant interest in polytomous IRT models only began in the 1980s. Polytomous models differ from dichotomous ones in that test items are not merely scored as right or wrong; instead, each response category is evaluated and scored according to its degree of correctness or the amount of information it contributes toward a full answer (partial credit model).

In the context of NOUN, where students hail from diverse socio-economic backgrounds, age groups, and geographical regions, applying these theories presents both opportunities and challenges. Ability estimates generated from these models are more than just statistical outputs; they serve as crucial indicators of student learning outcomes and inform instructional practices and policy decisions at the institutional level (Okonkwo & Nwafor, 2019). Ensuring the accuracy and fairness of these estimates is particularly challenging in an ODL environment, where factors such as test delivery modes, technological infrastructure, and student preparedness significantly influence assessment outcomes (Adeoye, 2020).

Moreover, scaling item parameters and scoring models to accommodate NOUN's large student population requires continuous refinement and validation. The potential for differential item functioning (DIF), due to the diverse student demographics, demands rigorous testing to ensure that assessments remain equitable and valid across different student groups (Olawale & Fatokun, 2022). This underscores the need for ongoing research and the adaptation of assessment practices to better align with the specific needs of distance education students in Nigeria.

Thus, this study aims to explore the application of item parameters, scoring models, and ability estimates in the context of distance education at NOUN. By analyzing the effectiveness of these models and understanding their impact on student assessments, the research seeks to contribute to the broader discourse on enhancing assessment practices in ODL settings. The findings offer valuable insights into how NOUN can further refine its assessment strategies to ensure fairness, accuracy, and academic integrity in its educational offerings.

Objectives

Specifically, the study seeks to:

i. estimate the psychometric properties of mathematics tests developed based on 1-PL, 2-PL and 3-PL models.
ii. determine the difference between male and female students' mean ability scores in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored based on dichotomous and polytomous scoring models.

Research Questions

The researcher answered the following research questions:

i. What are the psychometric properties of mathematics tests developed based on 1-PL 2-PL and 3-PL item parameter models?
ii. What is the difference between male and female students' mean ability scores in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored based on dichotomous and polytomous scoring models?

Null Hypotheses

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

There is no significant difference between male and female students' mean ability scores in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored based on:

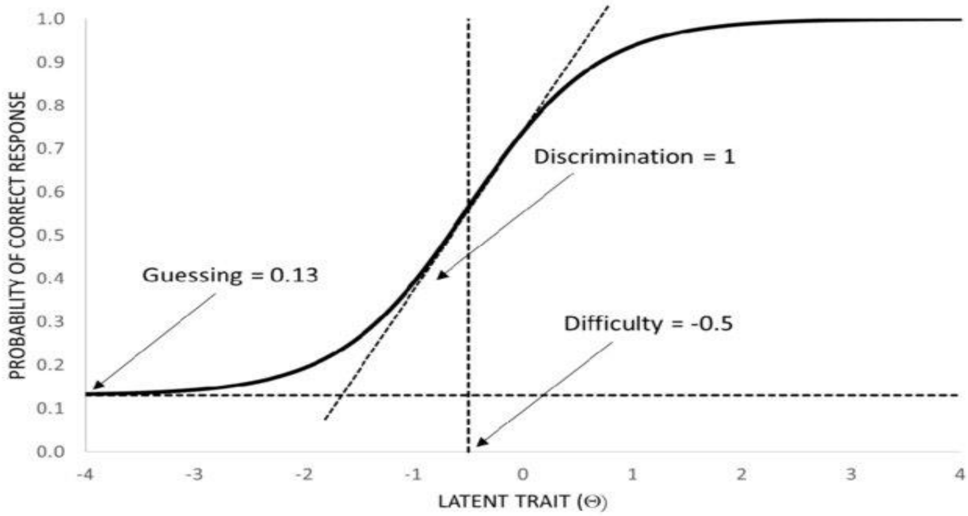
dichotomous and
polytomous scoring models.

Theoretical Framework

Classical Test Theory (CTT) is one of the most widely used frameworks for understanding the reliability and validity of test scores. According to CTT, an observed test score is the sum of a true score (representing the test taker's actual ability) and an error score (accounting for random fluctuations in performance). The theory assumes that all test items contribute equally to the total score and that measurement error is normally distributed and uncorrelated with true scores (Obidike & Adewale, 2021). In the context of distance education, CTT is often applied to evaluate the reliability and validity of test items used in assessments. For example, Tella and Bashorun (2019) highlight the importance of using CTT to assess the reliability of online assessments in Nigerian universities, including NOUN. They emphasize that while CTT provides valuable insights, its limitations in handling multidimensional data and its assumption of equal item contribution can sometimes reduce its effectiveness in diverse educational settings.

Item Response Theory (IRT) represents a more sophisticated approach to modeling the relationship between a student's latent ability and their performance on test items. Unlike CTT, IRT assumes that different items can have varying levels of difficulty and discrimination, allowing for a more nuanced analysis of test performance. IRT is particularly useful in distance education contexts where student populations are diverse and ability levels can vary widely (Olumuyiwa, 2019). Studies have shown that IRT provides a more accurate estimate of student ability by accounting for the variability in item characteristics. Igbokwe and Ogili (2020) argue that IRT models, such as the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL), offer better insights into student ability estimates in Nigerian distance education programs. Research conducted at NOUN demonstrated that IRT-based assessments could improve the precision of ability estimates and inform better instructional strategies.

Figure 1: An Item Characteristic Curve in IRT



Source: Belov (2011)

Conceptual Framework

Concept of Item Parameters in IRT Context: Aderinoye and Ojokheta (2018) assert that Item Response Theory (IRT) attempts to model the relationship between an unobserved variable, typically conceptualized as an examinee's ability, and the probability of the examinee correctly responding to a given test item. Three commonly used models in IRT are the 3-parameter logistic, the 2-parameter logistic, and the 1-parameter logistic models, with the 1-parameter model often referred to as the Rasch Model. These models all assume a single underlying ability, typically denoted as β, which is a continuous, unbounded variable. However, they differ in the characteristics they assign to test items. Each model incorporates an item difficulty parameter (b), which represents the point of inflection on the ability (β) scale. For both the 1-parameter and 2-parameter models, this is the point on the β scale where an examinee has a 50% probability of answering the item correctly. In the 3-parameter model, the probability of answering correctly at this point is (1 + c)/2, where c is the lower asymptote parameter, often referred to as the guessing parameter.

Theoretically, difficulty values (b) can range from -∞ to +∞, though in practice, they usually fall between -3 and +3, when scaled to a mean of 0 and a standard deviation of 1.0. Similarly, examinee ability values (β) can also range from -∞ to +∞, but values exceeding ±3 are rarely encountered. Items with high b values are considered difficult, as low-ability examinees have a lower probability of answering them correctly, whereas items with low b values are easier, allowing even low-ability examinees a moderate chance of providing the correct response (Ayala, 2009).

In addition to the difficulty parameter, the 2- and 3-parameter models include a discrimination parameter (a), which enables items to distinguish between examinees with varying levels of ability. Technically, the a parameter represents the slope of the item characteristic curve (ICC) at the point of inflection. The a value can range from -∞ to +∞, with typical values around 2.0 for multiple-choice items. Higher a values indicate that an item discriminates more effectively between examinees near the point of inflection. The 3-parameter model also incorporates a lower asymptote parameter (c), often referred to as the pseudo-guessing parameter, which accounts for the likelihood of low-ability examinees guessing the correct answer.

Scoring Models in Educational Assessment: Polytomous Scoring Model: Item response model for items with more than two response categories, e.g. multiple-choice item that allows partial credits for each of the response categories, or constructed-response item with multiple steps (Belov, 2011). Dichotomous Scoring Model: Item response model for test with binary items. Examinees taking the test will respond in either one of the two response categories. A test with items scored right or wrong is dichotomous (Okonkwo & Nwafor, 2019).

Components of psychometric analysis are as follows:

i. **Reliability**: According to Igbokwe and Ogili (2020), it is the consistency of a test in measuring what it is intended to measure.
ii. **Validity**: The extent to which a test measures what it is supposed to measure.

Item Analysis:

i. **Difficulty Index**: Measures the proportion of students who answer a test item correctly. It helps in understanding how challenging a question is (Cook & Foster, 2012).
ii. **Discrimination Index**: Assesses how well a test item differentiates between high and low performers. High discrimination indices indicate that a question is effective at distinguishing between different levels of student ability.
iii. **Distractor Analysis**: Evaluates the effectiveness of distractors (incorrect answer choices) in multiple-choice questions. Effective distractors should be plausible to students who do not know the correct answer (Adeoye, 2020).

Test Structure and Cognitive Levels:

i. **Blueprinting**: Ensures that test items are distributed appropriately across various cognitive levels and content areas as per the curriculum.
ii. **Bloom's Taxonomy**: Analyzes the cognitive levels of questions to ensure a balanced assessment (Alabi, 2021).

Fairness and Bias Analysis:

i. **Bias Detection**: Identifies and addresses any potential biases in test items that could unfairly advantage or disadvantage certain groups of students.
ii. **Equity Analysis**: Ensures that all students, regardless of background or ability, have an equal opportunity to perform well on the test (Kpoloyie, 2017).

Methodology

The study utilized an evaluative survey research design, with a sample consisting of 126 second-year mathematics students from the Faculty of Sciences at the National Open University of Nigeria (NOUN). The sample was evenly split, comprising 63 boys and 63 girls, selected through a stratified random sampling technique. Three achievement tests, based on the 1, 2, and 3 Parameter Logistic Models, were developed in line with the course content for MTH 211 (Abstract Algebra). The first instrument, referred to as the 1 Parameter Logistic Model - Mathematics Achievement Test (1PLM-MAT), the second as the 2 Parameter Logistic Model - Mathematics Achievement Test (2PLM-MAT), and the third as the 3 Parameter Logistic Model - Mathematics Achievement Test (3PLM-MAT), were all used to assess students' performance.

The instruments demonstrated logical validity indices of 0.80, 0.76, and 0.86 for 1PLM-MAT, 2PLM-MAT, and 3PLM-MAT, respectively. Additionally, the reliability coefficients obtained for the three tests were 0.82, 0.88, and 0.86, reflecting strong internal consistency. After administering the tests to the students, both dichotomous and polytomous scoring models were applied to score the tests, and the scores were collected for further analysis.

Psychometric analysis was conducted using the R package, which provides functions for reliability analysis, factor analysis, and descriptive statistics. It also focuses on Item Response Theory (IRT) models, including 1PL, 2PL, and 3PL. The Statistical Package for the Social Sciences (SPSS) was employed for Analysis of Variance (ANOVA) to determine statistical significance. To explore the data further, the Scheffe method was used for post-hoc comparisons, identifying the specific sources of significant differences.

Results

Research Question 1: What are the psychometric properties of mathematics tests developed based on 1-PL 2-PL and 3-PL models?

Table 1: Psychometric Properties of Tests Developed Based on 1-PL 2-PL and 3-PL Models

Test	Validity Index		Reliability Index		Difficulty Index		Discrimination Index		Distracter Index	
	OV	AR	OV	AR	OV	AR	OV	AR	OV	AR
1-PL Model										
Mathematics Test	0.84	≥ 0.7	0.82	≥ 0.7	0.46	.4 - .6	0.49	≥0.3	-.40	≥ - .4
2-PL Model										
Mathematics Test	0.80	≥ 0.7	0.84	≥ 0.7	0.42	.4 - .6	0.48	≥0.3	-.48	≥ - .4
3-PL Model										
Mathematics Test	0.80	≥ 0.7	0.72	≥ 0.7	0.47	.4 - .6	0.42	≥0.3	-.42	≥ - .4

Source: Field research, (2024), Note: OV-Value Obtained, AR-Acceptable Range

Table 1 presents the psychometric properties of the mathematics tests developed based on the 1-Parameter Logistic (1-PL), 2-Parameter Logistic (2-PL), and 3-Parameter Logistic (3-PL) models. A detailed analysis of the obtained values compared to acceptable ranges for all psychometric indices indicates that the three categories of tests meet the standards for being appropriate tools for ability estimation. According to Alabi (2021), the acceptable range for validity and reliability is between 0.7 and 1.0, the difficulty index should fall between 0.4 and 0.6, the discrimination index should range from 0.3 to 1.0, and the distractor index should be ≥ -0.4. All tests developed within this study align with these benchmarks, confirming their suitability for effective ability estimation.

Research Question 2: What is the difference between male and female students' mean ability scores in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored based on dichotomous and polytomous scoring models?

Table 2: Descriptive Statistics for Students' Ability in Mathematics Test Designed Based on 1-PL, 2-PL and 3-PL Models and Scored Based on Dichotomous and Polytomous Scoring Models

Item Parameter Model	Count	Mean				Standard Deviation			
		Scored Dichotomously		Scored Polytomously		Scored Dichotomously		Scored Polytomously	
		Male	Female	Male	Female	Male	Female	Male	Female
1-PL Model	126	30.9	29.9	30.9	29.9	12.1	9.6	12.1	9.6
2-PL Model		43.7	43.8	60.3	54.7	20.9	19.5	23.7	21.6
3-PL Model		44.5	41.5	61.6	58.9	21.5	19.7	26.9	23.2

Source: Field research, (2024)

Table 2 presents the descriptive statistics comparing the mean ability scores of male and female students in mathematics tests designed based on the 1-Parameter Logistic (1-PL), 2-Parameter Logistic (2-PL), and 3-Parameter Logistic (3-PL) models, and scored using dichotomous and polytomous scoring models. The results indicate that both male and female students demonstrated similar academic ability in the 1-PL model, as shown by the relatively close mean and standard deviation values for both groups, regardless of whether the dichotomous or polytomous scoring model was used. However, the data reveal that both male and female students achieved higher scores when the 2-PL and 3-PL models were paired with the polytomous scoring model. This suggests that students of both genders performed their best in mathematics when tests based on the 3-PL model were scored polytomously, as evidenced by the highest mean and standard deviation values, highlighted in the shaded areas of Table 2.

Hypothesis 1: There is no significant difference between male and female students' mean ability scores in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored based on:

dichotomous and
polytomous scoring models.

Table 3: F-test Statistics for Difference in Students' Mean Scores in Mathematics Tests Designed Based on 1-PL, 2-PL and 3-PL Models and Scored Based on Dichotomous Model

Source of Variation	**Sum of Squares**	**Df**	**Mean Square**	**F_{cal.}**	***P_{value}***	**Remark**
Male Students						
Between Sets	2417.342	2	150.444	2.406	3.000	Not Significant
Within Sets	73.394	63	62.543			
Total	2490.736					
Female Students						
Between Sets	1958.570	2	100.576	2.362	3.000	Not Significant
Within Sets	75.222	63	42.586			
Total	2033.792					

Source: Field research, (2024)

Table 3 presents the F-test statistics for the significant difference between male and female students' mean ability scores in mathematics tests designed based on the 1-Parameter Logistic (1-PL), 2-Parameter Logistic (2-PL), and 3-Parameter Logistic (3-PL) models, all scored using the dichotomous scoring model. The results show that both male and female students' achievements, at the 0.05 level of significance with 2 degrees of freedom for between sets and infinity degrees of freedom for within sets, yielded F-test values of 2.406 and 2.362, respectively. These values are less than the critical value of 3.000. As a result, since the calculated F-test values are below the critical threshold, the null hypothesis is accepted, and the alternative hypothesis is rejected. This implies that both male and female students demonstrated similar (and relatively low) achievement in mathematics tests designed based on the 1-PL, 2-PL, and 3-PL models when scored dichotomously.

Table 4: F-test Statistics for Difference in Students' Mean Scores in Mathematics Tests Designed Based on 1-PL, 2-PL and 3-PL Models and Scored Based on Polytomous Model

Source of Variation	**Sum of Squares**	**Df**	**Mean Square**	**$F_{cal.}$**	**$F_{tab.}$**	**Remark**
Male Students						
Between Sets	3958.385	2	301.564	6.780	3.000	Significant
Within Sets	93.920	63	44.475			
Total	4052.305					
Female Students						
Between Sets	1856.999	2	181.222	4.695	3.000	Significant
Within Sets	76.394	63	38.598			
Total	1933.393					
Post-hoc Comparism (Test after F-test)						
1-PL Model	Vs		2-PL Model			Significant
1-PL Model	Vs		3-PL Model			Significant
2-PL Model	Vs		3-PL Model			Not Significant

Source: Field research, (2024),

Table 4 shows the F-test statistics for the significant difference between male and female students' mean ability scores in mathematics tests designed based on the 1-Parameter Logistic (1-PL), 2-Parameter Logistic (2-PL), and 3-Parameter Logistic (3-PL) models, scored using the dichotomous scoring model. The results reveal that at a 0.05 level of significance, with 2 degrees of freedom for between sets and infinity degrees of freedom for within sets, the F-test values are 6.780 and 4.695, respectively, both of which exceed the critical value of 3.000. Therefore, since the calculated F-test values are higher than the critical value, the null hypothesis is rejected. Post-hoc analysis indicates a significant difference between students' achievements in tests designed based on the 1-PL model compared to those based on the 2-PL and 3-PL models. However, there is no significant difference between the achievements in tests designed based on the 2-PL and 3-PL models. This suggests that both male and female students demonstrated similar (and relatively high) achievement in mathematics tests designed based on the 2-PL and 3-PL models when scored polytomously.

Findings

Both male and female students demonstrated relatively low achievement in mathematics tests designed based on 1-PL, 2-PL and 3-PL models and scored dichotomously.

Again, both male and female students demonstrated relatively high achievement in mathematics tests designed based on 2-PL and 3-PL models and scored polytomously. However, the test designed based on 1-PL model and scored polytomously was found to produce low estimate of students' ability.

The post-hoc analysis revealed that there was a significant difference between students' achievement tests designed based on 1-PL and 2-PL/3-PL models. But no significant difference between students' achievement tests designed based on 2-PL and 3-PL models and scored polytomously. Therefore, implying a better estimate of students' ability when tests designed based on 2-PL and 3-PL models and scored polytomously.

Conclusion

Based on the findings of this study, the following conclusions can be drawn: Polytomous scoring models offered more reliable ability estimates for students than dichotomous scoring models under the 2-Parameter Logistic (2-PL) and 3-Parameter Logistic (3-PL) models. This finding highlights a key distinction between Classical Test Theory (CTT) and Item Response Theory (IRT) regarding the treatment of measurement error, as indicated by the standard error of measurement. Furthermore, the study emphasizes the crucial role of item parameters and scoring models in accurately assessing distance education students at the National Open University of Nigeria (NOUN). To improve the quality of assessments, it is essential for NOUN to adopt more sophisticated scoring models and to continually review item parameters. These enhancements will not only yield more precise evaluations of student abilities but also foster more effective and equitable educational outcomes in distance learning environments.

Recommendations

Based on the findings of this study, the following recommendations were offered:

1. Based on the large difference in ability estimation that this study has found between dichotomous scoring and polytomous scoring models, the polytomous scoring model which provides for partial credit be embraced to improve on the accuracy of ability estimation in school-based assessment. Since test scores are used to guide educational decisions therefore, the accuracy of their outcomes should be a priority.
2. The NOUN should consider incorporating more advanced psychometric models, such as Item Response Theory (IRT) and Computerized Adaptive Testing (CAT), to improve the accuracy and reliability of ability estimates. These models can better account for the diverse abilities of distance education students and provide more tailored assessment outcomes.
3. Regularly calibrate and review test items to ensure they remain valid and reliable across different cohorts of students. This process involves analyzing item parameters such as difficulty, discrimination, and guessing to maintain the integrity of assessments and ensure they accurately measure student abilities.
4. NOUN should explore and potentially integrate a variety of scoring models, including those that incorporate partial credit scoring and polytomous item responses. This diversification would allow for a more nuanced assessment of student performance, especially in disciplines where binary scoring may not fully capture student learning outcomes.
5. Invest in robust data analytics tools that can handle large datasets, allowing for more sophisticated analysis of item parameters and student ability estimates.

This capability will enable NOUN to identify trends, detect anomalies, and continuously improve the assessment process.

References

Adeoye, B. F. (2020). Exploring the challenges and opportunities in online learning at the National Open University of Nigeria. *African Educational Research Journal,* **8** (2): 423-432. DOI: 10.4018/978-1-5225-9746-9

Aderinoye, R. A., & Ojokheta, K. O. (2018). Challenges of implementing item response theory in open and distance learning assessments. *African Journal of Educational Assessment,* **9** (1): 45-61.

Alabi, A. T. (2021). The role of item response theory in the development of standardized tests: A case study of the National Open University of Nigeria. *Journal of Educational Assessment in Africa,* **15** (3): 102-114.

Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford Press.

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement,* **35**: 495-517.

Cook, R. J., & Foster, C. C. (2012). *Application of the 3PL IRT model to the detection of shifterrors.* Paper presented at the annual meeting of the Northeastern Educational Research Association. Rocky Hill, CT.

Igbokwe, B. C., & Ogili, C. A. (2020). Application of Item Response Theory in estimating student ability in distance education: A case study of the National Open University of Nigeria. *Journal of Educational Measurement and Evaluation*, **16** (2): 125-138.

Kpolovie, P. J. (2017). *Test, measurement, and evaluation in education*. Port Harcourt: University of Port Harcourt Press.

Obidike, C. A., & Adewale, S. O. (2021). Using Generalizability Theory to assess the reliability of student scores in distance education: A case study at the National Open University of Nigeria. *African Journal of Educational Research and Development*, **23** (1): 89-102.

Okonkwo, C. A., & Nwafor, C. E. (2019). Assessment in open and distance learning in Nigeria: Issues and challenges. *International Journal of Educational Technology and Learning,* **6** (2): 53-59.

Olawale, A. R., & Fatokun, K. V. (2022). Addressing differential item functioning in computer based testing: Evidence from the National Open University of Nigeria. *Journal of Educational Measurement in Africa,* **11** (4): 87-101.

Olumuyiwa, T. A. (2019). Assessment in distance education: Analyzing the reliability and validity of test items at the National Open University of Nigeria. *International Journal of Distance Education and E-Learning,* **7** (3): 112-129.

Tella, A., & Bashorun, M. T. (2019). Reliability and validity of online assessments: Classical Test Theory (CTT) perspective. *International Journal of Educational Technology*, **14** (3): 97-109.