

**ASSESSMENT OF DIFFERENTIAL ITEM FUNCTIONING OF BECE 2022 BASIC SCIENCE AND TECHNOLOGY ACHIEVEMENT TEST OF UPPER BASIC III STUDENTS IN BENUE STATE, NIGERIA**

<sup>1</sup> OPITA Matthew John,  
*opitamatthewjohn@gmail.com*

<sup>2</sup>A.D.E Obinne & <sup>3</sup>S.O. Emaikwu  
<sup>1,2,3</sup> Department of Guidance and Counselling, Joseph Sarwuan Tarka University, Makurdi,

**Article Info**

*Article History:*

*Received: 25, Mar., 2026*

*Revised: 01, Apr., 2026*

*Accepted: 30, Apr., 2026*

*Published: 05 May., 2026*

**Keywords:**

Differential Item  
Functioning,  
Dimensionality,  
Item fit,  
Achievement Test,  
BECE

**Abstract**

This study investigated Differential Item Functioning (DIF) in the BECE 2022 Basic Science and Technology Multiple Choice Achievement Test among Upper Basic III students in Zone C of Benue State, Nigeria. The study was motivated by persistent performance disparities across demographic groups, raising concerns about the fairness and validity of the assessment. An ex-post facto research design was adopted. The population comprised 16,062 candidates, from which a sample of 3,142 students with complete response data was purposively selected. Data were analysed using Item Response Theory (IRT) procedures, including dimensionality analysis with NOHARM, item fit analysis using the S-X<sup>2</sup> statistic under the three-parameter logistic (3PL) model in JMetrik, and Differential Item Functioning analysis using the Cochran-Mantel-Haenszel (CMH) method. Findings revealed that the test is multidimensional, with six underlying factors, thereby violating the unidimensionality assumption of standard IRT models. Item fit analysis showed that 80% of the items misfit the 3PL model, indicating substantial model-data incompatibility. Location based DIF was substantial, affecting 70% of items. The study concludes that the BECE 2022 Basic Science test exhibits significant psychometric limitations in terms of dimensionality, model fit, and fairness. It recommends the integration of DIF analysis into test development, the review of biased items, and the adoption of more, measurement models to enhance the validity and equity of large-scale assessments.

**How to Cite:**

Opita, M. J., Obinne, A.D.E., & Emaikwu, S.O. (2026). Assessment of Differential Item Functioning of BECE 2022 Basic Science and Technology Achievement Test of Upper Basic III Students in Benue State, Nigeria. *International Journal of Contemporary Education and Management*. 1(1), 75–86

## INTRODUCTION

Basic Science and Technology (BST), often referred to as Integrated Science, constitutes the foundational science subject for students at the upper basic level of education in Nigeria. It is designed to introduce learners to fundamental scientific concepts that underpin advanced disciplines such as Biology, Chemistry, and Physics. As emphasized in the National Policy on Education, BST plays a critical role in promoting scientific literacy and technological development necessary for national growth.

The curriculum developed by the Nigerian Educational Research and Development Council outlines key objectives of BST, including fostering students' interest in science, developing problem-solving and critical thinking skills, and enabling the application of scientific knowledge to real-life situations. The realization of these objectives depends largely on effective instruction and, importantly, the quality of assessment practices used to evaluate student learning.

Assessment serves as a systematic process for collecting and interpreting information about students' achievement. In science education, it is essential for monitoring progress, identifying learning gaps, and guiding instructional decisions. However, the validity of assessment outcomes is contingent upon the quality and fairness of the measurement instruments. One major threat to test fairness is Differential Item Functioning (DIF), which occurs when test items function differently for groups of examinees with equal ability.

DIF represents a critical concern in educational measurement because it introduces systematic bias into test scores. When present, it may advantage or disadvantage certain groups—such as male or female students, or students from urban or rural schools— independent of their actual ability. This undermines the validity of score interpretations and may lead to inequitable educational decisions.

Large-scale assessments such as the Basic Education Certificate Examination (BECE) rely heavily on multiple-choice test formats due to their efficiency, objectivity, and broad content coverage. Despite these advantages, multiple-choice items are not immune to bias, making it necessary to evaluate their fairness through rigorous psychometric procedures such as Item Response Theory (IRT) and DIF analysis.

Preliminary evidence from BECE results in Zone C of Benue State indicates persistent performance disparities across gender. While these differences may reflect genuine variations in ability, they may also suggest the presence of biased test items. Without empirical investigation, it remains unclear whether such disparities are attributable to true differences in achievement or measurement artifacts.

Given the high-stakes nature of BECE and its role in determining students' academic progression, ensuring the fairness and validity of its test items is imperative. DIF analysis provides a robust framework for identifying potentially biased items and improving the quality of assessment instruments.

### **Statement of the Problem**

Despite the importance of fairness in standardized assessments, there is limited empirical evidence on whether BECE Basic Science and Technology test items function equivalently across different student groups in Zone C of Benue State. Observed disparities in performance across gender raise concerns about possible item bias.

If such bias exists and remains undetected, it may distort students' true achievement levels, perpetuate educational inequalities, and compromise the credibility of the BECE as a fair assessment tool. Therefore, a systematic investigation of Differential Item Functioning in the BECE 2022 Basic Science and Technology Achievement Test is necessary.

### **Objective of the study:**

The study investigated Differential Item Functioning of BECE 2022 Basic Science Achievement Test of Upper Basic III Students in Zone C of Benue State, Nigeria. The specific objectives of the study include, to:

1. determine the dimensionality of the BECE 2022 Basic Science and Technology Multiple Choice Test in Zone C of Benue State, Nigeria.

2. find out the item fit of the BECE 2022 Basic Science and Technology Multiple Choice Test in Zone C of Benue State, Nigeria.
3. identify the items in the BECE 2022 Basic Science and Technology Multiple Choice Test that function differentially based on gender in Zone C of Benue State, Nigeria.

### **Research Questions**

The study sought answers to the following research questions:

1. What are the Dimensionality of the BECE 2022 Basic Science and Technology Multiple Choice Test in Zone C of Benue State, Nigeria?
2. What is the item fit of BECE 2022 Basic Science and Technology Multiple Choice Test in Zone C of Benue State Nigeria?
3. How many items in the BECE 2022 Basic Science and Technology Multiple Choice Test function differentially based on gender in Zone C of Benue state Nigeria?

### **METHODOLOGY**

#### **Research Design:**

This study adopted an ex post facto research design, which was considered appropriate because the data analyzed were pre-existing and the independent variable gender was not subject to manipulation. The design enabled the examination of differential item functioning (DIF) across naturally occurring groups without experimental intervention.

The study was conducted in Zone C of Benue State, Nigeria, a region characterized by both urban and rural settings and a mix of co-educational and single-sex schools. This heterogeneity provides a suitable context for investigating DIF across diverse educational environments. The zone comprises several Local Government Areas, including Otukpo, Ohimini, Ogbadibo, Ado, Okpokwu, Agatu, Obi, Oju, and Apa. Zone C was selected due to its diversity in school types and locations, as well as preliminary evidence of performance disparities in BECE results across gender. These characteristics make the zone particularly appropriate for examining potential sources of item bias.

The population of the study consisted of all Upper Basic III (JSS 3) students who sat for the 2022 Basic Education Certificate Examination (BECE) in Basic Science in Zone C of Benue State. According to official records from the Benue State Examinations Board, the population comprised 16,062 students. A sample of 3,129 students was drawn from the population using purposive sampling. Only students with complete response data across all 60 test items and complete demographic information (gender) was included. Records that were incomplete or illegible were excluded.

The sample size was considered adequate for Item Response Theory (IRT) and DIF analyses, which require large samples for stable parameter estimation. The selected sample was subsequently categorized based on gender. The instrument for this study was the BECE 2022 Basic Science and Technology multiple-choice test, obtained from the Benue State Examinations Board. The dataset consisted of students' responses to 60 multiple-choice items, along with demographic information. Responses were scored dichotomously, with correct answers coded as 1 and incorrect or omitted responses coded as 0. This coding procedure is consistent with standard IRT practices and enabled the construction of a binary response matrix used for subsequent analyses. The instrument had been previously validated by the examination body during test development using standard procedures. Reliability was established using Cronbach's alpha, with a coefficient of 0.80, indicating acceptable internal consistency. Data were obtained from the Benue State Examinations Board upon official approval. The researcher collected students' response data and the corresponding marking scheme. All responses were coded and organized into a structured dataset for analysis. Data were analyzed using Item Response Theory (IRT) procedures with the aid of NOHARM and JMetrik statistical software.

Dimensionality was assessed using linear factor analysis in NOHARM, employing the Goodness-of-Fit Index (GFI) and Root Mean Square Residual (RMSR). A GFI value of .90 or higher and a reduction in RMSR of less than 10% were used as criteria for determining unidimensionality. Item fit was evaluated using the S-X<sup>2</sup> statistic under the three-parameter logistic (3PL) model in JMetrik. Items with  $p > .05$  were considered to have acceptable fit, while those with  $p \leq .05$  were classified as misfitting. Differential Item Functioning (DIF) was examined using the Cochran–Mantel–Haenszel (CMH) method across gender. Items were classified based on the Educational Testing Service (ETS) guidelines into Category A (negligible DIF), Category B (moderate DIF), and Category C (large DIF), using both statistical significance and effect size.

## Presentation of Results

**Research Question One:** What is the dimensionality of the BECE 2022 Basic Science Multiple Choice test in Zone C of Benue State, Nigeria?

To answer this research question, linear factor analysis with NOHARM software was used.

**Table 1:** linear factor analysis with NOHARM results on Dimensionality of BECE 2022 Basic Science Test

BECE Test	Items	No. of Dimension	GFI	RMSR	% Reduction in RMSR
BECE 2022 Basic Science	60	6 (F1, F2, F3, F4, F5, F6)	0.8895391	0.047634	% < 10 (8.5%) at 7 HD

*Note:*  $F$  = Factors underlying a test,  $HD$  = Number of Hypothesized Dimension,  $GFI$  = Goodness of Fit Index,  $RMSR$  = Root Mean Square Residual.

**GFI:** ranges from 0 to 1;  $\geq .90$  = acceptable fit

**RMSR:** ranges from 0 upward; lower = better; < 10% reduction = parsimonious

Table 2 shows the dimensionality of the 60-item BECE 2022 Basic Science test administered to 3,142 candidates. The analysis revealed that the BECE 2022 Basic Science test has six dimensions (F1, F2, F3, F4, F5, and F6), with a Goodness of Fit Index ( $GFI$ ) = 0.8895391, which approaches the 90% benchmark, and a Root Mean Square Residual ( $RMSR$ ) = 0.047634. The six-dimensional model was parsimonious after calculating the percentage reduction in  $RMSR$ ; the value was less than 10% (8.5%) at 7 hypothesized dimensions. This implies that the BECE 2022 Basic Science test is multidimensional, with six factors (F1 through F6) underlying the test. This means the test violated the assumption of unidimensionality, and six factors accounted for the variation observed in candidates' performance on the BECE 2022 Basic Science test.

**Research Question Two:** What is the item fit of the BECE 2022 Basic Science Multiple Choice Test in Zone C of Benue State, Nigeria?

To answer this research question, **S-X<sup>2</sup> Statistic** in jmetrik was used.

**Table 2: Item Fit Statistics for the BECE 2022 Basic Science Test (3PL Model) using S-X<sup>2</sup> Statistic**

Item	S-X <sup>2</sup>	df	p-value	Decision	Item	S-X <sup>2</sup>	Df	p-value	Decision
qs1	155.3973	32	.000	Misfit	qs31	139.3813	31	.000	Misfit
qs2	85.4889	32	.000	Misfit	qs32	66.7607	31	.000	Misfit

qs3	48.7598	32	.029	Misfit	qs33	34.9107	32	.331	<b>Fit</b>
qs4	56.3925	31	.004	Misfit	qs34	98.6809	33	.000	Misfit
qs5	42.3985	31	.083	<b>Fit</b>	qs35	44.4199	32	.071	<b>Fit</b>
qs6	56.1599	31	.004	Misfit	qs36	44.5063	32	.070	<b>Fit</b>
qs7	149.5088	31	.000	Misfit	qs37	83.4970	31	.000	Misfit
qs8	39.5982	31	.138	<b>Fit</b>	qs38	35.1010	31	.280	<b>Fit</b>
qs9	70.0367	31	.000	Misfit	qs39	133.0422	32	.000	Misfit
qs10	69.3513	31	.000	Misfit	qs40	36.8740	32	.254	<b>Fit</b>
qs11	52.3162	32	.013	Misfit	qs41	52.5563	31	.009	Misfit
qs12	70.3942	31	.000	Misfit	qs42	63.6456	31	.001	Misfit
qs13	97.7363	32	.000	Misfit	qs43	31.1098	31	.461	<b>Fit</b>
qs14	58.8112	32	.003	Misfit	qs44	33.8845	32	.377	<b>Fit</b>
qs15	100.9551	31	.000	Misfit	qs45	83.8177	32	.000	Misfit
qs16	138.3790	30	.000	Misfit	qs46	61.5339	32	.001	Misfit
qs17	69.2595	32	.000	Misfit	qs47	123.4982	32	.000	Misfit
qs18	64.5657	32	.001	Misfit	qs48	127.0837	31	.000	Misfit
qs19	68.7975	32	.000	Misfit	qs49	57.0344	32	.004	Misfit
qs20	97.6384	31	.000	Misfit	qs50	59.3471	32	.002	Misfit
qs21	48.6633	32	.030	Misfit	qs51	52.9264	32	.011	Misfit
qs22	48.7431	31	.022	Misfit	qs52	60.4486	32	.002	Misfit
qs23	45.2399	33	.076	<b>Fit</b>	qs53	46.3226	32	.049	Misfit
qs24	87.1953	32	.000	Misfit	qs54	70.2961	32	.000	Misfit
qs25	44.1044	32	.075	<b>Fit</b>	qs55	72.2364	32	.000	Misfit
qs26	49.5132	32	.025	Misfit	qs56	82.1675	31	.000	Misfit
qs27	79.9559	31	.000	Misfit	qs57	63.8596	32	.001	Misfit
qs28	44.7410	31	.053	<b>Fit</b>	qs58	128.9532	32	.000	Misfit
qs29	75.9363	32	.000	Misfit	qs59	103.1466	32	.000	Misfit
qs30	65.5685	30	.000	Misfit	qs60	76.3690	33	.000	Misfit

---

Note: Item fit:  $p > .05 = fit$ ;  $p \leq .05 = misfit$

Item fit was assessed using the  $S-X^2$  statistic (Orlando & Thissen, 2000). Items with  $p > .05$  are considered to fit the 3PL model. Items shown in bold under Decision column indicate fitting items. Items with adequate fit ( $p > 0.05$ ): 12 items (20.0%), Items with poor fit ( $p < 0.05$ ): 48 items (80.0%)

Table 2 presents the item fit statistics for all 60 items of the BECE 2022 Basic Science test under the three-parameter logistic (3PL) IRT model. Item fit was evaluated using the  $S-X^2$  statistic developed by Orlando and Thissen (2000), which compares observed and expected item response frequencies across ability score groups. An item is considered to fit the model when the  $S-X^2$  p-value exceeds the .05 significance level, indicating that the difference between observed and model-expected responses is not statistically significant. Out of 60 items, only 12 items (qs5, qs8, qs23, qs25, qs28, qs33, qs35, qs36, qs38, qs40, qs43, and qs44) demonstrated acceptable fit to the 3PL model ( $p > .05$ ), representing 20.0% of the total items. The remaining 48 items (80.0%) showed statistically significant misfit ( $p \leq .05$ ), indicating that the observed response patterns for these items deviated substantially from what the 3PL model predicted. This high rate of item misfit suggests that while the 3PL model provides a reasonable overall framework for the test as indicated by the model-level fit indices in Table 3, a considerable proportion of individual items do not conform adequately to the model's assumptions. These misfitting items may require further review in terms of their content, scoring, or psychometric properties before the test can be considered fully calibrated under the 3PL IRT framework.

**Research Question Three:** How many items in the BECE 2022 Basic Science Multiple Choice Test function differentially based on location in Zone C of Benue State, Nigeria?

To answer this research question, **Cochran-Mantel-Haenszel DIF** in jmetrik was used.

Table 3 Presents the Cochran-Mantel-Haenszel DIF analysis results for location groups (Urban as reference group, Rural as focal group) based.

Item	Chi-square	p-value	Effect Size	95% C.I.	DIF Class	Direction
qs1	111.58	0.00	-5.85	(-4.45, -7.25)	C-	Favors Rural
qs5	11.17	0.00	-2.28	(-0.85, -3.70)	B-	Favors Rural
qs6	37.50	0.00	2.65	(3.53, 1.77)	C+	Favors Urban
qs7	133.91	0.00	5.08	(6.13, 4.03)	C+	Favors Urban
qs8	15.57	0.00	1.66	(2.51, 0.81)	B+	Favors Urban
qs9	42.62	0.00	2.64	(3.47, 1.81)	C+	Favors Urban
qs11	53.87	0.00	3.07	(3.96, 2.19)	C+	Favors Urban
qs12	13.42	0.00	-1.76	(-0.77, -2.74)	B-	Favors Rural
qs15	31.83	0.00	2.59	(3.53, 1.64)	C+	Favors Urban
qs16	33.87	0.00	-2.50	(-1.61, -3.39)	C-	Favors Rural
qs18	15.26	0.00	1.66	(2.49, 0.82)	B+	Favors Urban
qs19	57.50	0.00	3.04	(3.87, 2.22)	C+	Favors Urban

qs20	93.53	0.00	-3.76	(-2.92, -4.59)	C-	Favors Rural
qs21	12.79	0.00	1.66	(2.61, 0.72)	B+	Favors Urban
qs22	35.47	0.00	-2.21	(-1.45, -2.98)	C-	Favors Rural
qs25	6.81	0.01	-1.15	(-0.27, -2.03)	B-	Favors Rural
qs26	31.49	0.00	2.18	(2.99, 1.38)	C+	Favors Urban
qs28	11.32	0.00	-1.47	(-0.61, -2.33)	B-	Favors Rural
qs29	39.22	0.00	2.68	(3.55, 1.80)	C+	Favors Urban
qs30	26.83	0.00	2.19	(3.04, 1.33)	C+	Favors Urban
qs31	4.71	0.03	1.14	(2.16, 0.11)	B+	Favors Urban
qs32	77.46	0.00	-3.69	(-2.79, -4.60)	C-	Favors Rural
qs33	28.86	0.00	-2.25	(-1.41, -3.08)	C-	Favors Rural
qs34	7.43	0.01	1.21	(2.12, 0.31)	B+	Favors Urban
qs35	80.07	0.00	-3.79	(-2.90, -4.68)	C-	Favors Rural
qs36	61.84	0.00	-3.41	(-2.55, -4.27)	C-	Favors Rural
qs37	28.35	0.00	-2.33	(-1.44, -3.23)	C-	Favors Rural
qs38	25.74	0.00	-2.26	(-1.35, -3.17)	C-	Favors Rural
qs39	73.82	0.00	-3.28	(-2.46, -4.10)	C-	Favors Rural
qs40	94.49	0.00	-4.99	(-3.79, -6.18)	C-	Favors Rural
qs41	6.75	0.01	-1.15	(-0.28, -2.03)	B-	Favors Rural
qs44	9.62	0.00	1.33	(2.23, 0.44)	B+	Favors Urban
qs46	16.19	0.00	-1.70	(-0.85, -2.56)	B-	Favors Rural
qs47	69.12	0.00	3.70	(4.66, 2.75)	C+	Favors Urban
qs48	6.65	0.01	1.04	(1.85, 0.22)	B+	Favors Urban
qs50	67.08	0.00	4.05	(5.10, 2.99)	C+	Favors Urban
qs51	28.47	0.00	-2.18	(-1.36, -3.00)	C-	Favors Rural
qs52	35.39	0.00	2.57	(3.43, 1.70)	C+	Favors Urban
qs54	20.08	0.00	-2.05	(-1.16, -2.95)	C-	Favors Rural
qs55	6.95	0.01	1.12	(1.96, 0.27)	B+	Favors Urban

qs56	48.95	0.00	-3.05	(-2.15, -3.95)	C-	Favors Rural
qs58	29.67	0.00	2.00	(2.80, 1.21)	C+	Favors Urban
qs59	24.15	0.00	1.80	(2.54, 1.05)	C+	Favors Urban

**Note:**  $\Delta MH$  = Delta Mantel-Haenszel effect size. A positive  $\Delta MH$  value indicates DIF favouring the reference group; a negative value indicates DIF favouring the focal group  
 Category: A (Negligible) Effect Size  $|\Delta MH| < 1.0$   $p$ -value Criterion Not significant ( $p > .05$ ) Decision: No DIF item retained,  
 B (Moderate)  $1.0 \leq |\Delta MH| < 1.5$  Significant ( $p \leq .05$ ) Decision: Slight to moderate DIF flagged for review  
 C (Large)  $|\Delta MH| \geq 1.5$  Significant ( $p \leq .05$ ) Decision: Moderate to large DIF substantial bias

The location-based DIF analysis revealed a substantial level of differential item functioning based on school location. A striking 42 out of 60 items (70%) exhibited significant DIF between urban and rural students. Of these, 25 items favored rural students while 17 items favored urban students. Thirty items showed moderate to large DIF (Category C), indicating substantial bias. The magnitude of some effect sizes was extremely large, such as item qs1 (ES = -5.85 favoring rural), item qs7 (ES = 5.08 favoring urban), and item qs40 (ES = -4.99 favoring rural). This finding suggests serious equity concerns in the test, as a majority of items functioned differently for students from urban versus rural educational settings. This bias could reflect differences in curriculum exposure, teaching quality, resource availability, or cultural contexts between urban and rural schools that may advantage one group over the other beyond their actual Basic Science proficiency.

## DISCUSSION

The present study investigated the psychometric properties and fairness of the BECE 2022 Basic Science and Technology Achievement Test using dimensionality analysis, item fit statistics, and Differential Item Functioning (DIF) procedures. The findings provide important insights into both the structural validity and equity of the assessment.

The dimensionality analysis revealed that the test is multidimensional, with six latent factors underlying item responses. This finding indicates a violation of the unidimensionality assumption required for conventional Item Response Theory (IRT) models such as the 3PL. The multidimensional structure is, however, not entirely unexpected given the integrative nature of Basic Science and Technology, which draws from diverse domains such as physics, chemistry, biology, and technology. Similar patterns of multidimensionality have been reported in large-scale science and mathematics assessments, particularly those designed to sample broad curricular content. This suggests that the construct being measured is inherently complex and may not be adequately represented by a single latent trait.

The implication of this finding is critical: applying a unidimensional IRT model to a multidimensional test may lead to biased parameter estimates and compromised model fit, which was evident in the present study. The item fit analysis confirmed this concern, as a substantial proportion of items (80%) exhibited significant misfit under the 3PL model. Such a high rate of misfit suggests a mismatch between the model assumptions and the empirical data. It is likely that the multidimensional nature of the test, combined with potential issues in item construction (e.g., ambiguous wording or ineffective distractors), contributed to this outcome. This finding aligns with previous studies which have shown that more complex models do not necessarily yield better fit when underlying assumptions are violated. Consequently, the use of simpler or multidimensional IRT models may be more appropriate for such data.

Beyond structural validity, the study examined test fairness through DIF analysis across gender. The results revealed varying degrees of measurement inequity across school type (co-educational vs. single-sex) was found to be a substantial source of DIF, with more than half of the items functioning differentially. This finding highlights the influence of institutional context on student performance. Differences in instructional approaches, peer interactions, and academic culture between school types may shape how students interpret and respond to test items. The large effect sizes observed for several items suggest that some items may be measuring context-specific experiences rather than general science ability. This poses a threat to score comparability and calls for greater attention to contextual fairness in test development.

## Conclusion

Based on the findings, the following conclusions were drawn.

The BECE 2022 Basic Science Achievement Test exhibited multidimensionality with six underlying dimensions, indicating that the test does not measure a single underlying construct as required by standard IRT assumptions. The item fit analysis using the S-X<sup>2</sup> statistic revealed that 48 items (80%) exhibited significant misfit to the 3PL model, with only 12 items (20%) demonstrating adequate fit. This high rate of item misfit suggests that the majority of items in the test did not conform well to the expectations of the 3PL IRT model, largely due to the test's multidimensional structure.

Furthermore, evidence of differential item functioning (DIF) was identified across school location, indicating the presence of measurement bias in a substantial proportion of test items. The Location-based bias was the most severe, affecting 43 items (71.7%). Consequently, the study concludes that the BECE 2022 Basic Science Achievement Test administered to Upper Basic III students in Zone C of Benue State, Nigeria, contains items that function differentially across examinee subgroups, thereby raising serious concerns about the validity, fairness, and comparability of score interpretations particularly across students from different geographical locations and school types.

## Recommendations

1. Benue state Examination body responsible for developing and administering the BECE should incorporate differential item functioning analysis as a mandatory step in the test development process.
2. The examination body should establish a standing Fairness Review Panel that includes educators, subject matter experts, and assessment specialists with diverse backgrounds representing urban and rural schools. This panel should review all test items for potential sources of bias related to content, context, language, and cultural references that might differentially advantage or disadvantage specific subgroups.

## References

- Adamoah, Y. K., & Acquah, J. (2016). Determinants of Students' Performance in Basic Education Certificate Examination (BECE) in the Upper East Region of Ghana: A Case Study of Kassena-Nankana West District (Knwd). *American Journal of Research and Communication*, 4(10) 91-107.
- Adeoye, F. A. (2018). Challenges of standardizing educational assessment in Nigeria. *Journal of Educational Assessment in Africa*, 7(2), 24-32.
- Adewale, O. (2020). *Educational assessment in Nigeria: Issues and challenges*. Ibadan: Spectrum Books.
- Ajayi, K., & Adeniyi, K. (2020). Regional disparities in educational achievement in Nigeria: Policy implications. *African Educational Review*, 18(3), 189-210.
- Ajeigbe, T. O., & Afolabi, E. R. I. (2014). Assessing Unidimensionality and Differential Item Functioning in Qualifying Examination for Senior Secondary School Students, Osun State, Nigeria,4 (4),30 – 37

- Alade, O.M, Aletan, S & Sokenu, B.S. (2020). Assessing the differential item functioning of 2018 WASSCE Mathematics Achievement Tests in Lagos State, Nigeria. *African Journal of Behavioural and Scale Development Research AJB-SDR*, 2(2), 8-20
- Alonge, M. F. (2017). *Evaluation in Education: A Practical Guide*. Ibadan: Stirling-Horden Publishers.
- Anikweze, C. M. (2013). *Measurement and Evaluation for Teacher Education*. Enugu: Snaap Press.
- Ayanwale, M. A, Isaac-Oloniyo, F. O, & Abayomi, F. R. (2020). Dimensionality assessment of binary response test items: A non-parametric approach of Bayesian item response theory measurement. *International Journal of Evaluation and Research in Education (IJERE)*, 9 (2), 385- 393.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443-459.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3), 261-280.
- Bolt, D. M., & Stout, W. F. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviour metrika*, 23(1), 67-95.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Chalmers, R. P. (2018)**. Model-based measures for detecting and quantifying response bias. *Psychometrika*, 83(3), 696-732.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31-44.
- De Ayala, R. J. (2019). *The theory and practice of item response theory*. New York: Guilford Press.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Lawrence Erlbaum Associates.
- Drabinová, A., & Martinková, P. (2017)**. Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, 54(4), 498-517.
- Effiom, A. P (2021). Test Fairness and Assessment of Differential Item Functioning of Mathematics Achievement Test for Senior Secondary Students in Cross River State, Nigeria. *Global Journal of Educational research*, 20 (1),55-62.
- Egwu, O. (2015). The alignment of curriculum content and achievement tests: A case study of Basic Science in Nigeria. *Educational Measurement in Africa*, 4(2), 38-45.

- Emaikwu, S. O. (2019) *Fundamentals of Research Methodology and statistics*. (Revised edition) Makurdi: Selfers Academic Press Publishers
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Evans, G. U., Uko M. P. & Ekim, R. E. D (2022) Investigation of Differential Item Functioning of Basic Education Certificate Examination (BECE) Mathematics Items in Akwa Ibom State, Nigeria. *The African Journal of Behavioural and Scale Development Research AJB-SDR* (4) 1
- Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.
- Faremi, Y. A., & Jimoh, S. A. (2022). Differential item functioning and implications for testing in Nigeria education system. *Indonesian Journal of Learning Education and Counseling*, 5(1), 1-10.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359-374.
- Frisbie, D. A. (2018). Reliability of multiple-choice tests. In J. P. Keeves (Ed.), *Educational measurement* (pp. 242-255). Pergamon Press.
- Haladyna, T. M., & Downing, S. M. (2019). *Validity of multiple-choice tests: Construct validity and content-related evidence*. *Educational and Psychological Measurement*, 60(5), 812-822.
- Haladyna, T. M., & Rodriguez, M. C. (2018). *Developing and validating test items* (4th ed.). New York: Routledge.
- Hambleton, R. K. (2018). Good practices for identifying differential item functioning. *Medical Care*, 44(11), S182-S188.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Lawrence Erlbaum Associates.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Ibrahim, A. (2018). Differential item functioning: The state of the art. *Jigawa Journal of Multidisciplinary Studies (JJMS)*, 1(1), 37-50.
- Igbokwe, D. O. (2019). Administration of national examinations in Nigeria: Problems and prospects. *Journal of Educational Management*, 12(1), 56-73.
- Ige, A. M. (2021). Evolution of educational assessment policies in Nigeria: Historical perspectives. *International Journal of Educational Research*, 9(1), 52-68.
- Isiugo-Abanihe, I.M., Ugwoke, M. & Iluobe, I. O. (2018). An Empirical Investigation of Goodness-of-fit of 1-, 2- and 3- Parameter Logistic Models to an Achievement Test Data in English Language. *Journal of Educational Assessment in Africa*. 13(1), 14-24

Jimoh, Abdulaziz & Oniye (2020) Assessment of Items Prone to Guessing in SSCE Economics Multiple-Choice tests among Students in Kwara State, Nigeria. *Anatolian Journal of Education*. (5)1 pp. 17-28