International Journal of Academic Pedagogical Research (IJAPR)
ISSN: 2643-9123
Vol. 6 Issue 9, September - 2022, Pages: 36-48

Test Models and Dimensionality Assessment of Mathematics Constructed Response Items

Babatunde Kasim Oladele[1], Olutayo Toyin Omole[2], Owotorufa Ebi Obiyo[3]

Institute of Education, University of Ibadan, Nigeria
oladelebatunde@gmail.com
Institute of Education, University of Ibadan, Nigeria
vicayoade@gmail.com
Ministry of Education, Yenagoa, Bayelsa
ebijosephine@gmail.com

Abstract: *Test items from public examination bodies are expected to be valid and reliable, particularly the constructed-response items aspect of such examinations, which measures more than one trait of the test takers. The dimensionality of Mathematics constructed-response test items from two prominent Nigerian public examinations was investigated in this study. The examination bodies are the West African Examinations Council (WAEC) and the National Examinations Council (NECO). To evaluate the dimensionality of the two test forms, two relevant test models were compared. A descriptive survey design was used. The population of this study included all senior secondary school three (SSS3) students in Ibadan Metropolis in Oyo state, Nigeria. To select 24 secondary schools, a simple random sampling technique was used, and 1151 students participated in this study. For the instruments for data collection, three years of compiled past questions from the two examinations (2013-2015) were used. Classical test and item response theory models were used to analyze the data. Under the classical test theory (CTT), explanatory factor analysis, scree plot, and parallel analysis were used to determine dimensionality, whereas in the item response theory, generalized partial credit and graded response model (IRT) were used. When the CTT models were used, the results revealed that WAEC constructed-response items have four dimensions, whereas NECO constructed items have three dimensions. WAEC and NECO both have three dimensions under the IRT models. It was determined that the constructed-response test items of the two examination bodies are multidimensional and that the IRT model provided better estimates than the CTT.*

Keywords: Public examinations, Test models and dimensionality, Mathemathetics Constructed-Response items

1. INTRODUCTION

Public examination bodies saddled with the responsibility of producing test items are expected to have standardized and quality items. The need to have valid and reliable test items calls for a determination of such test dimensionality. The interaction between items and examinees is reflected in dimensionality. All dimensions of the tests could be unidimensional or multidimensional. When a single trait can explain or account for examinee test performance, a set of test items is considered unidimensional; when more than one trait can explain or account for examinee test performance, a set of test items is considered multidimensional. When only one dominant dimension is detected, data are considered essentially unidimensional; when more than one dominant dimension is detected, data are considered multidimensional (Zhang, 2016). Furr and Bacharach (2013), on the other hand, defined general dimensionality assessment as (a) determining the number of dimensions of test items, (b) estimating the relationship between compound dimensions (if any), and (c) mapping statistical dimensions to psychological attributes. One of the fundamental assumptions of test theory is that a score can only have meaning if the set of items measures only one attribute or dimension. Over time, educational assessment experts have confirmed that dimensionality is closely related to equating, a statistical procedure that allows for the comparison of test scores from various forms (Kolen & Brennan, 2014).

Dimensionality is the process of acquiring precise knowledge of a test's internal structure and a good understanding of how test marks established human abilities through dimensionality assessment. Some dimensionality assessment outcomes allow test authors and users to thoroughly validate detailed explanations and applications of test scores (Zhang, 20016). Nontrivial dimensions are those that are significantly related to more than five items (Stone & Yeh, 2006). Score reliability and dimensionality are related. Dimensionality is a type of test score, and as such, the scoring method used to generate those scores has an impact. To score item responses, for example, a special issue could be considered. The number of dimensionalities may influence the reliability coefficient. As a result, it appears critical to examine the extent to which test forms and examinee dimension structures are characterised. The number of examinees included in the analysis for each test form is critical for dimensionality assessment. When a sample is drawn at random from the population, regardless of sample size, a dimensionality assessment method should yield roughly the same results.

To assess test dimensionality, various methods, such as exploratory factor analysis (EFA), linear factor analysis, nonparametric tests for essential unidimensionality, and the use of multidimensional IRT models, are used. Stone and Yeh (2006) state that one advantage

of IRT models over linear factor analytic methods is that information from examinee response patterns is analysed rather than the more limited information from correlation matrices. According to Embretson and Reise (2000), multidimensional IRT models have been used to assess the dimensionality of tests with items representing various skills, knowledge, and understanding. Eigenvalue examination is a widely used method in both the EFA and IRT literature for determining the number of latent traits in the model (Zopluoglu & Davenport, 2017). Previous research has revealed the Kaiser-Guttman rule (KG; Guttman, 1954; Kaiser, 1960), as well as the subjective scree test (Cattell, 1966). The three most commonly used eigenvalue examination methods in the EFA literature for determining the number of latent traits is parallel analysis (Horn, 1965; Green et al., 2012). The Kaiser-Guttmann rule is based on variance accounting. A complete system of non-redundant items has a variance equal to the number of items. To be stronger than average, a dimension must have an eigenvalue greater than one (Zopluoglu and Davenport, 2017). The KG rule then declares "significant" all dimensions with eigenvalues greater than one. The scree test is a plot of the eigenvalues of the principal component versus its dimension. Eigenvalues are larger than the average for "real" dimensions; however, after a point (dimension), all eigenvalues decrease uniformly because the remaining variance accounted for by the eigenvalues is random. The slope of the eigenvalue to dimension becomes constant when the eigenvalues decrease uniformly. From now on, the plot will be nothing more than a line (Zopluoglu and Davenport Jr, 2017). It is the investigator's responsibility to determine (subjectively) when the scree plot transforms into a line separating significant and random dimensions.

Dimensionality assessment methods can be divided into two families based on the statistical procedure used: parametric and nonparametric. Nonparametric methods require fewer or stronger models and assumptions than parametric methods. It is deceptive to make hasty connections between the type of procedure (parametric or nonparametric) and the definition of dimensionality (EFA or IRT) while ignoring the fundamental difference between parametric and nonparametric methods. Some EFA-based techniques are parametric, while others are not. A good example is Buja and Eyuboglu's (1992) parallel analysis for EFA (PA). In contrast to the original PA (Horn, 1965), which is a parametric procedure, their version is nonparametric by substituting the permutation principle for the normality assumption. Although some nonparametric methods based on conditional covariances, such as DIMTEST 10 (Nandakumar and Stout, 1993; Stout, 1987), have been widely used in IRT, Miller and Hirsch's (1992) method appears parametric due to its reliance on MIRT models. Frequently, dissimilar dimensional solutions are obtained. All of these variables have an impact on the performance of dimensionality assessment methods, both directly and indirectly. Different methods frequently produce dissimilar dimensional solutions, but no method has ever been shown to be universally satisfactory (Stone and Yeh, 2006; Svetina and Levy, 2014; Tate, 2003; Beltz Verlag van Abswoude, van der Ark, and Sijtsma, 2004). However, the default methods of some popular software packages may significantly distort one's understanding of dimensionality. Researchers and practitioners must weigh potential gains and losses when choosing specific methods.

When using item response theory frameworks, the dimensionality of the test items is one of the most important factors to consider. The number of latent variables recognized by a test is defined as its dimensionality. All unidimensional tests focus on estimating a single latent variable, whereas multidimensional tests focus on estimating multiple latent variables. Selected-response test items are typically unidimensional, whereas constructed-response test items are by definition multidimensional. This is because in answering a Mathematics CR test item, the examinee is expected to demonstrate some traits, such as computational skills, the skill to decode problems into mathematical language and demonstrate logical or abstract thinking. As stated in both WAEC and NECO syllabi, the Mathematics Examination aims to examine candidates (Nigerian Educational Research and Development Council, 2013): (i) mathematical and computational ability, (ii) understanding of mathematical concepts and their relationships to the acquisition of entrepreneurial skills for daily living in the global world, (iii) the ability to translate problems into mathematical language and solve them using appropriate methods, (iv) accuracy relevant to the problem at hand, and (v) logical, abstract, and precise thinking. Tests are one of the best tools for measuring and assessing students' abilities in the educational system (Oladele & Adegoke, 2020). The main objective of this study is to identify the dimensions of WAEC and NECO CR test items. The scoring process chosen frequently influences how various dimensionality assessment methods act and what results are obtained. The widespread use of a CR item, in contrast, entails students' constructed-response. Various forms of constructed-response tests have been seen in active tests, such as short answers, essays, and speaking prompts. The constructed-response items are typically scored on an integer scale of 0 to 5, according to various predetermined scoring rubrics. In the scoring of constructed-response tests, both manual and engine raters are used; for tests that use multiple raters, the number of raters and approaches for ratings vary across dissimilar items.

In Nigeria, the majority of candidates received low scores on Mathematics constructed-response items in public examinations, affecting their final grade on the examination. Most public examination bodies are not used to evaluating the dimensionality of their test items, particularly constructed-response test items. As a result, this study investigates the dimensionality status of WAEC and NECO mathematics to determine whether these test items are unidimensional or multidimensional. The current investigation aims to present measures to address problems of low scores encountered by candidates in external examinations such as the WAEC and NECO. For this purpose, answers were provided to the following research questions:

1. What is the minimum dimensionality of WAEC Mathematics constructed-response items among senior secondary school students in Ibadan Metropolis using:

i. Classical test theory models?
ii. Item response theory models?

2. What is the minimum dimensionality of NECO Mathematics constructed-response items among senior secondary school students in Ibadan Metropolis using:
    i. Classical test theory models?
    ii. Item response theory models?
3. How comparable are the established minimum dimensionality of WEAC and NECO under:
    i. Classical test theory models?
    ii. Item response theory models?

2. METHOD

The study used a descriptive survey research design. All senior secondary school students in Ibadan Metropolis were included in the study's population. where the sample was drawn. From the 24 senior secondary schools chosen at random, 1151 students (male =565, female =586) participated. The students' ages ranged from 12 to 16 (46.7 percent), 17 to 19 (52.4 percent), and 20 to 22 (52.4 percent) (1.0 percent). Past constructed-response test items from WAEC and NECO for three years (2013-2015) were used as data collection instruments. Before the instruments were used in the schools, the Oyo State Government, school administration, and students all gave their approval. The data collected were analyzed with classical test models (exploratory factor analysis, scree plot test and parallel analysis) and item response models (generalized partial credit and graded response models).

3. FINDINGS

3.1 Research question 1: What is the minimum dimensionality of WAEC mathematics constructed-response items among senior secondary school students in the Ibadan Metropolis using:
    i. Classical test theory models?
    ii. Item response theory models?

To provide an answer to research question 1, the test items' sample adequacy was determined and is presented in Table 1.

Table 1: KMO and Bartlett's Test Statistics of 15 WAEC Mathematics Constructed Achievement Test Items

	Criterion	Value
	KMO	0.65
Bartlett's Test of Sphericity	Approx. Chi-Square	2209.52
	df	105
	p-value	0.00

Table 1 shows that the statistic obtained for KMO sampling adequacy was 0.65, which was relatively good. The statistics of Chi-Square for Bartlett's Test of Sphericity were significant at a p-value < 0.05. This implies that the test data were adequate and followed a normal distribution.
Thus, exploratory factor analysis was performed on test data to determine the number of dimensions.

Table 2: Total Variance (WAEC Mathematics Constructed-Response Achievement Test)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	2.55	16.99	16.99	2.55	16.99	16.99
2	1.77	11.82	28.81	1.77	11.82	28.81
3	1.54	10.26	39.08	1.54	10.26	39.08
4	1.24	8.26	47.34	1.24	8.26	47.34
5	1.10	7.32	54.65	1.10	7.32	54.65
6	1.01	6.73	61.38	1.01	6.73	61.38
7	0.86	5.74	67.13			
8	0.80	5.34	72.46			
9	0.76	5.03	77.50			
10	0.74	4.92	82.42			

11	0.63	4.18	86.60
12	0.54	3.62	90.22
13	0.54	3.57	93.79
14	0.49	3.25	97.04
15	0.44	2.96	100.00

Table 2 presents the statistics of the dimensionalities of the WAEC constructed-response items. The total variance explained output was used, as revealed in Table 2, and the highest eigenvalue was 2.55 for component one. This indicated that the highest component explained was 16.99% with an eigenvalue of 2.55. The acceptable rule is that extracted factors when put together explaining 50% to 60% of the variance with eigenvalues greater than one should be kept as good extracted values (Oladele, 2021). This implies that the test data have six underlying factors. The six factors showed that if the test data have more than one dimension, then the test data may be considered approximately multidimensional. This indicates that the WAEC Mathematics constructed-response item is multidimensional with six dimensions. In addition, a scree plot was also constructed to further confirm the dimensionality of the test.

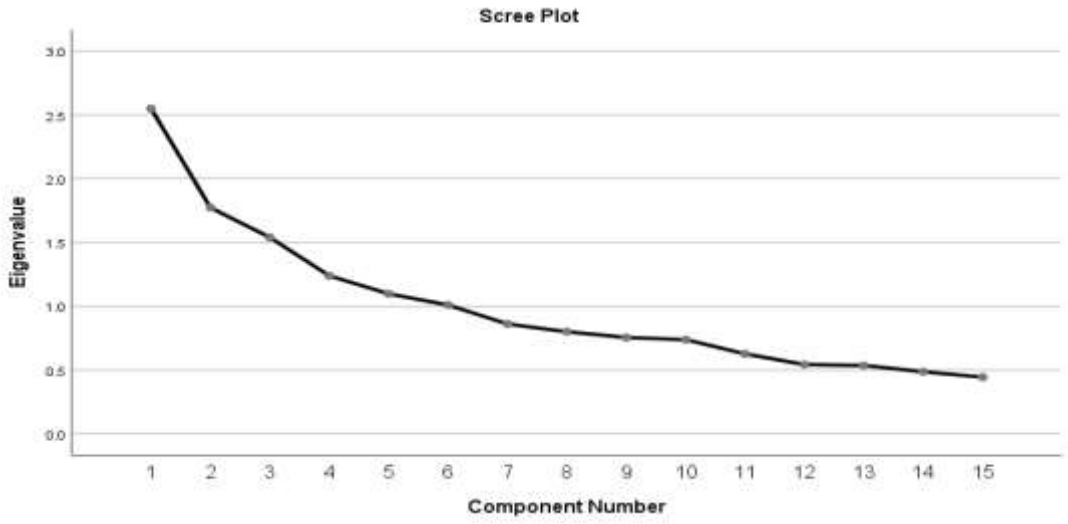

Figure 1: Scree plot for WAEC Mathematics Constructed-response Test Items

Figure 1 shows that the components on the y-axis move down towards the X-axis. The dots represent this downwards slope in terms of their contribution to the variance. Every space between two dots represents a factor. The rule of thumb is to keep all components or factors within the sharp descent before the eigenvalues trail off (Steven, 2012; Adegoke, 2013). As a result, the number of factors underlying the test data is greater than one, indicating multidimensionality. In addition, additional analysis was carried out to validate the dimensionality of the test data. Parallel analysis was performed using Monte Carlo Principal Component Analysis software. The generation of a set of random correlation matrices is required for parallel analysis. After subjecting these random correlation matrices to principal component analysis, the average of their eigenvalues is computed and compared to the eigenvalues produced by the experimental data (Watkins, 2006). This is presented in Table 3.

Table 3: Comparison of Monte Carlo PCA for Parallel Analysis Statistics for 15 Adopted WAEC Mathematics Test Items

Items	Real Data Eigenvalues	Randomly Generated Data Eigenvalues	Standard Deviation
1	2.55*	1.20	0.023
2	1.77*	1.27	0.018
3	1.54*	1.16	0.014
4	1.24*	1.21	0.012
5	1.10	1.12	0.012
6	1.01	1.17	0.010
7	0.86	1.09	0.009
8	0.80	1.12	0.011
9	0.76	1.07	0.011
10	0.74	1.09	0.012
11	0.63	1.04	0.011
12	0.54	1.05	0.011
13	0.54	1.02	0.014
14	0.49	1.02	0.014
15	0.44	1.00	0.819

*Suggested dimensions: 4

Table 3 shows the comparison of the eigenvalues (experimental and generated data), while there were four components of the real data set with eigenvalues (2.55, 1.77, 1.54, and 1.24) greater than the eigenvalues (1.31, 1.27, 1.24, and 1.21) of the generated data set. The result implies that there are likely four factors that underlie the performance of examinees in WAEC constructed response items. This also suggests that WAEC constructed response items are multidimensional with a minimum of four dimensions and consequently measured four traits. Furthermore, exploratory factor analysis was conducted again based on the minimum number of dimensions suggested by the Parallel Analysis Statistics (PAS) to identify items that measure the objectives of WAEC Mathematics. The results of the analysis are presented in Table 4.

Table 4: Rotated Factor Matrix of WAEC Constructed-response Items

Factor				
Item	**1**	**2**	**3**	4

1	**0.547***	**0.408***	0.138	0.078
2	**0.651***	-0.01	-0.041	-0.390
3	**0.655***	-0.149	-0.23	-0.131
4	**0.703***	0.057	0.05	0.132
5	**0.511***	0.348	0.228	-0.245
6	-0.021	0.114	**0.805***	-0.016
7	0.117	0.342	-0.095	**-0.490***
8	-0.007	0.017	-**0.748***	-0.075
9	-0.074	**0.588***	0.007	0.075
10	-0.185	**0.562***	-0.263	-0.226
11	0.115	**0.714***	0.078	-0.038
12	0.179	**0.457***	0.136	0.015
13	**0.442***	-0.019	-0.272	**0.545***
14	-0.051	-0.211	0.189	0.202
15	-0.129	0.162	0.090	**0.732***

*Absolute loading values > 0.4

Table 4 shows the four factors and item loadings on them. The varimax rotation method with Kaiser Normalization, which is an orthogonal rotation technique, was applied. Thus, Table 4 shows that items 1, 2, 3, 4, 5, and 13 are highly loaded on factor one, items 1, 9, 10, 11, and 12 are highly loaded on factor two, and items 6 and 8 are highly loaded on factor three, while items 7, 13 and 15 are highly loaded on factor four. However, it was only item 14 that did not load highly on any factor. This result shows that four substantial factors underlie WAEC constructed-response tests. Furthermore, the results show that items 1 and 13 loaded highly on more than one factor. For example, item 1 loaded highly on factors one and two, while item 13 loaded highly on factors one and four. This result suggests that WAEC constructed-response tests are multidimensional; that is, four abilities accounted for the observed variation in the performance of candidates in the test. Similarly, this suggests that the WAEC constructed-response test items measure more than one trait. Hence, there is a need for the examinees to possess much ability to provide the correct answer to the items. These mathematics abilities include factor one (mathematical competency and computational skills), factor two (understanding of mathematical concepts and their relationship to the acquisition of entrepreneurial skills for everyday living in the global world), factor three (translating problems into mathematical language and solving them using appropriate methods) and factor four (accurate to a degree relevant to the problem at hand and logical, abstract and precise thinking).

Furthermore, item response theory models (polytomous graded response model and generalized partial credit model of IRT-PRO Version 3.0) were used to determine the minimum level of the dimension of the WAEC mathematics constructed-response items. Table 5 shows the analysis results.

Table 5: Dimensionality of WAEC Mathematics Construction Response Test Using IRT (Generalized Partial Credit Model)

Dimension	Loglikelihood	Difference	p-value	Remark
1	48450.68			
2	47982.05	468.63		

3	47682.34	299.71	0.64*	Dimension limit

*p<1

Table 5 shows the minimum dimensionality of the level of WAEC test items using the generalized partial credit model (GPCM) as 4. This was determined based on the differences obtained when log likelihood values were compared, showing the p-values to be less than 1.

Table 6: Dimensionality of WAEC Mathematics Construction Response Test Using Graded Response Model (IRT)

Dimension	Loglikelihood	Difference	p value	Remark
1	48411.17			
2	47835.61	755.56		
3	47527.63	245.61	0.33*	Dimension limit

*p<1

Table 6 shows the minimum dimensionality of the level of WAEC test items using the graded response model (GRM) as 4. This was determined based on the differences obtained when log likelihood values were compared, showing the p-values to be less than 1.

3.2 Research Question 2: What is the minimum dimensionality of NECO mathematics constructed-response items among senior secondary school students in Ibadan Metropolis using:

iii. Classical test theory models?
iv. Item response theory models?

For the NECO test, the KMO test and Bartlett's test of sphericity were carried out to establish normality and sample adequacy. The results of the analysis for NECO mathematics constructed tests are presented in Table 7.

Table 7: KMO and Bartlett's Test Statistics of NECO Mathematics Constructed Achievement Test

	Criterion	Value
	KMO	0.72
Bartlett's Test of Sphericity	Approx. Chi-Square	2215.40
	Df	105
	p-value	0.00

Table 7 shows that the statistic obtained for KMO sampling adequacy was 0.72, which was relatively good. The statistics of Chi-Square for Bartlett's Test of Sphericity were significant at a p-value < 0.05. This implies that the test data were adequate and followed a normal distribution. Thus, exploratory factor analysis was performed on test data to determine the number of dimensions.

Table 8: Total Variance Explained (NECO Mathematics Constructed-Response) Achievement Test)

Component	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**		
S/N	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	2.80	18.69	18.69	2.80	18.69	18.69
2	1.85	12.35	31.04	1.85	12.35	31.04
3	1.41	9.40	40.44	1.41	9.40	40.44
4	1.07	7.15	47.59	1.07	7.15	47.59
5	1.03	6.84	54.43	1.03	6.84	54.43
6	0.92	6.14	60.57			
7	0.84	5.60	66.17			
8	0.77	5.15	71.32			
9	0.76	5.06	76.38			

10	0.71	4.72	81.10
11	0.69	4.57	85.67
12	0.62	4.15	89.82
13	0.55	3.66	93.48
14	0.52	3.49	96.97
15	0.45	3.03	100.00

To verify the number of dimensions of the NECO Mathematics constructed-response test items, the Explanatory Factor Analysis (EFA) was employed using SPSS version 25. Table 8 presents the statistics of the dimensionalities of the NECO constructed-response items. The total variance explained output was used, as revealed in Table 8, where the total variance explained by the highest eigenvalue was 2.80 for component one. This indicated that the highest component explained was 18.69% with an eigenvalue of 2.80. The acceptable rule is that extracted factors when put together explaining 50% to 60% of the variance with eigenvalues greater than one should be kept as good extracted values (Oladele, 2021). This implies that the test data have five underlying factors. The five factors showed that the test data have more than one dimension; thus, the test data may be considered approximately multidimensional. This indicates that the WAEC Mathematics constructed-response item is multidimensional with five dimensions. In addition, a scree plot was also constructed to further confirm the dimensionality of the test.

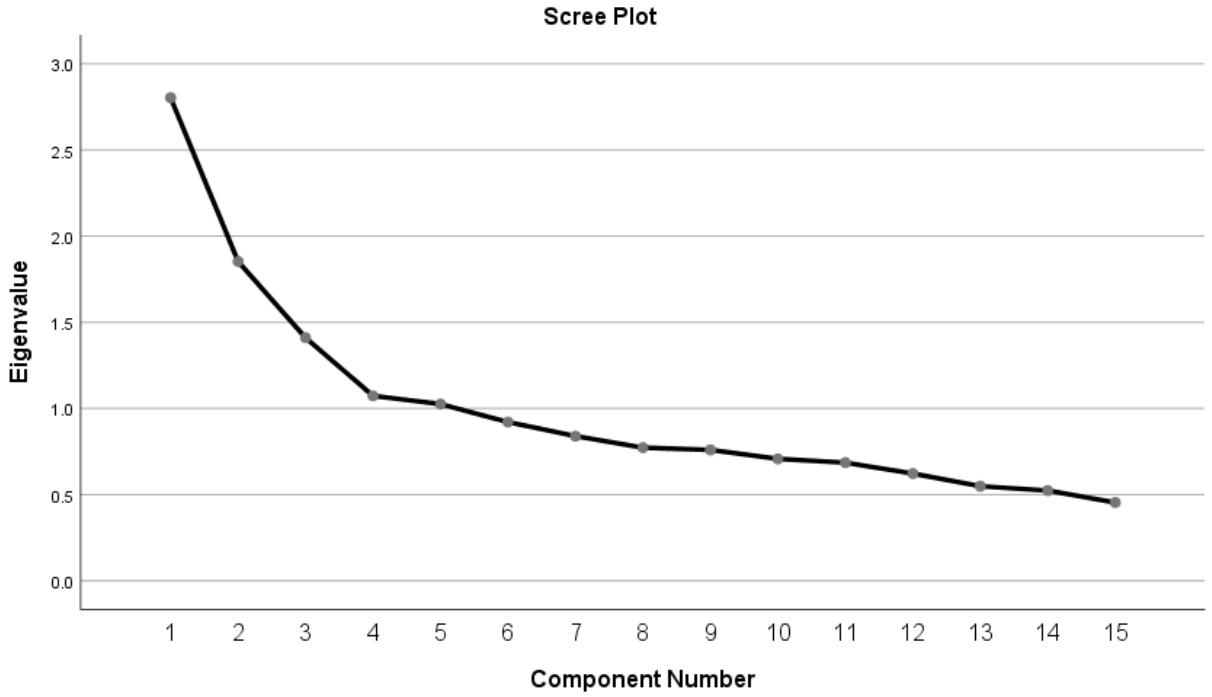

Figure 2: Scree plot of NECO mathematics constructed-response test items

Figure 2 shows that the components in the y-axis move down towards the X-axis. The dots represent the downwards slope in terms of their contribution to the variance. A factor is represented by each space between two dots. The acceptable rule is to keep all components or factors contained within the sharp descent before the eigenvalues trail off (Steven, 2012; Adegoke, 2013). As a result, there is more than one factor underlying the test data, indicating multidimensionality. Furthermore, additional analysis was performed to support the dimensionality of the test data. Monte Carlo Principal Component Analysis software was used for parallel analysis. For parallel analysis, a set of random correlation matrices with the same number of variables and respondents as the experimental data is needed. After that, the random correlation matrices are subjected to principal component analysis, with the average of their eigenvalues computed and compared to the eigenvalues of the experimental data. (Watkins, 2006). Table 9 displays the findings of the analysis.

Table 9: Monte Carlo PCA for Parallel Analysis Statistics for 15 Adopted NECO Mathematics Test

Items	Real Data Eigenvalues	Randomly Generated Data Eigenvalues	Standard Deviation
1	2.80*	1.20	0.023
2	1.85*	1.27	0.018
3	1.41*	1.16	0.014
4	1.07	1.21	0.012
5	1.03	1.12	0.012
6	0.92	1.17	0.010
7	0.84	1.09	0.009
8	0.77	1.12	0.011
9	0.76	1.07	0.011
10	0.71	1.09	0.012
11	0.69	1.04	0.011
12	0.62	1.05	0.011
13	0.55	1.02	0.014
14	0.52	1.02	0.014
15	0.45	1.00	0.819

*Suggested dimensions: 3

Table 9 shows the comparison of the eigenvalues (experimental and generated data), while there were four components of the real data set with eigenvalues (2.80, 1.85, and 1.41) greater than the eigenvalues (1.32, 1.28 and 1.25) of the generated data set. The result implies that there are likely three factors that underlie the performance of examinees in NECO constructed response items. This also suggests that NECO constructed response items are multidimensional with a minimum of three dimensions and consequently measured three traits. Furthermore, exploratory factor analysis was conducted again based on the minimum number of dimensions suggested by the Parallel Analysis Statistics (PAS) to identify items that measure the objectives of WAEC Mathematics. The results of the analysis are presented in Table 10.

Table 10: Rotated Factor Matrix of NECO Constructed-response Items

	Factor		
Item	**1**	**2**	**3**
1	0.388	0.124	0.052
2	**0.664***	-0.074	-0.254
3	**0.627***	-0.217	-0.278
4	**0.467***	0.175	-0.224
5	**0.519***	0.057	0.222
6	0.250	**0.586***	0.026
7	**0.483***	-0.140	-0.109
8	0.089	**0.515***	-0.046
9	**0.584***	0.353	0.100

10	0.316	**0.543***	0.154
11	**0.562***	0.210	0.264
12	-0.22	**0.527***	-0.278
13	-0.177	**0.534***	0.125
14	0.095	-0.066	**0.785***
15	-0.224	0.095	**0.743***

***absolute loading values > 0.4

Table 10 shows the three factors and item loadings on them. The varimax rotation method with Kaiser Normalization, which is an orthogonal rotation technique, was applied. Thus, Table 10 shows that items 2, 3, 4, 5, 7, 9 and 11 are highly loaded on factor one, Items 6, 8, 12 and 13 are highly loaded on factor two, while items 14 and 15 are highly loaded on factor three. However, it was only item 1 that did not load highly on any factor. This finding indicates that three important factors underpin NECO constructed-response tests. This finding suggests that NECO constructed-response tests are multidimensional, as three abilities accounted for the observed variation in test performance. Similarly, the NECO constructed-response test items appear to measure more than one trait. As a result, examinees must have more than one skill to provide correct answers to the items. One of these mathematical abilities is factor one (mathematical competency and computational skills), factor two (understanding of mathematical concepts and their relationship to the acquisition of entrepreneurial skills for everyday living in the global world), and factor three (translating problems into mathematical language and solving them using appropriate methods).

Item Response Theory Models (Polytomous Graded Response Model and Generalized Partial Credit Model of IRT-PRO Version 3.0) were used to determine the minimum level of the dimension of the NECO Mathematics constructed-response items. The results of the analysis are presented in Tables 11 and 12.

Table 11: Dimensionality of NECO Mathematics Construction Response Test Using IRT (Generalized Partial Credit Model)

Dimension	Log likelihood	Difference	p-value	Remark
1	51260.94			
2	50706.25	554.69		
3	50458.34	247.91	0.45*	Dimension limit

*p<1

Table 11 shows the minimum dimensionality of the level of NECO test items using the generalized partial credit model (GPCM) as 3. This was determined based on the differences obtained when log likelihood values were compared, showing the p-value (0.45) to be less than 1.

Table 12: Dimensionality of NECO Mathematics Construction Response Test Using IRT (Graded Response Model)

Dimension	**Log likelihood**	**Difference**	**p-value**	**Remark**
1	51307.31			
2	47835.61	347.17		
3	47527.66	307.95	0.89	Dimension limit

*p<1

Table 12 shows the minimum dimensionality of the level of NECO test items using the graded response model (GRM) as 3. This was determined based on the differences obtained when log likelihood values were compared, showing the p-value (0.89) to be less than 1.

3.3 Research Question 3: How comparable are the established minimum dimensionalities of WEAC and NECO under:

i. Classical test theory models?
ii. Item response theory models?

The comparability of the test models (classical test theory models and item response theory) is presented in Table 13 and Figure 1.

Table 13: Comparison of Test Dimensions under CTT and IRT

Test Model	Test Dimension		Remark
	WAEC	NECO	
CTT	4	3	Multidimensional
IRT (GPCM)	3	3	Multidimensional
IRT (GRM)	3	3	Multidimensional

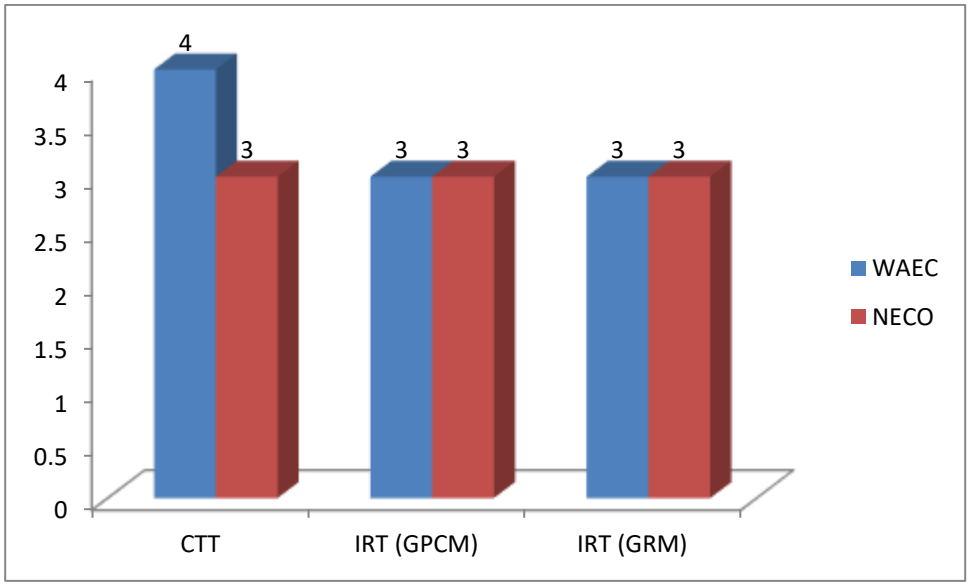

Table 13 and Figure 3 show that both the WAEC and NECO Mathematics constructed-response test items are multidimensional when compared under the classical test model and item response models (generalized partial credit model and graded response model). It was only the CTT model that revealed that WAEC test items have four dimensions, while the other models revealed that both WAEC and NECO Mathematics constructed-response test items have three dimensions.

4. DISCUSSION

The findings of this study showed that WAEC Mathematics constructed-response test items have four dimensions, while NECO Mathematics constructed-response test items have 3 dimensions based on the calibration with classical test models. On the other hand, it showed that WAEC and NECO Mathematics constructed-response test items have 3 dimensions based on the calibration with item response models. This implies that both test items are multidimensional and measure more than one trait. This is in line with the fact that CR tests require the students to demonstrate many abilities to be able to solve such mathematical problems that are couched in essay form. More importantly, WAEC and NECO syllabi show that students must be able to demonstrate critical thinking skill, the ability to manipulate data and the ability to properly present their work logically rather than just picking their answer from provided options in multiple-choice tests. The result supported Alu and Adediwura (2019), who conducted a dimensionality test on the 2015 and 2016 NECO Mathematics tests and concluded that the test items were multidimensional. This result also corroborated the findings of Metibemu (2020), who confirmed that WAEC Physics paper 1 has more than one dimension and measures more than one trait.

However, this result contradicts the findings of Ayanwale (2019), who established the 2015 NECO Mathematics constructed-response items to be unidimensional based on the framework used in establishing the number of dimensions of the test. From all these submissions, it could be inferred that most of the constructed-response test items of the public examinations are multidimensional and require the students to possess more than one latent trait for them to score good grades. Consequently, this could be one of the reasons why most of the students perform below average in mathematics. This gives credence to the findings of Reckase (1985) that some test items demand more than one latent trait or the ability to deal with, for instance, arithmetic and algebraic manipulations in mathematics.

These research findings revealed that both the WAEC and NECO Mathematics constructed-response test items are multidimensional. This in essence is for examination bodies to ensure that both their selected-response and constructed-response test items are unidimensional and not multidimensional. However, this is achievable by ensuring that the selection of the final test items processes is done by exposing the test items to a rigorous moderation system by test experts through qualitative measures with the view to enhancing the quality and reliability of these test items.

Acknowledgements: The researchers acknowledge and appreciate all participants (the students and schools) who served as samples of this study as well as my supervisor Professor B. A. Adegoke, who supervised my PhD thesis from which this article was extracted. I also acknowledge the financial subsidy given to me as a staff member by the University of Ibadan.Ibadan.

5. REFERENCES

Adegoke, B. A. (2013). Effects of the item-pattern scoring method on senior secondary school Student's ability scores in on the physics achievement test. *West African Journal of Education* 24: 181-190.

Ayanwale, M.A. (2019). Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous responses Mathematics achievement tests in Osun State, Nigeria *An unpublished PhD Thesis*, University of Ibadan.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. multivariate behavioural research, 27(4), 509-540.

Cattell, R. B. (1966). The scree test for the number of factors. multivariate behavioural research, 1(2), 245-276.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, 18, 41–68.

Embretson, S. E., and Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Furr, R. M., & Bacharach, V. R. (2013). Psychometrics: An introduction (2nd ed.). Thousand Oaks, CA: *SAGE Publications*, Inc.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W. (2012). A Proposed Solution to the problem with using completely random data to assess the number of factors with parallel analysis. Educational and Psychological Measurement, 72(3), 357-374.

Guttman, L. (1954). Some necessary and sufficient conditions for common factor analysis. *Psychometrika*, 19, 149-161.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, 9(2), 139-164.

Horn, J. L. (1965). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2), 179- 185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and psychological measurement, 20(1), 141-151.

Kolen, M. J., and Brennan, R. L. 2014. Test equating, scaling, and linking*: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag

MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, 39(2), 291-314.

Metibemu, M.A 2020. Assessment of the dimensionality and psychometric quality of the West African Examination Council (WASSCE) May/June 2014 Physics. In A, O. U. Onuka. *Public examining in Nigeria: A dynamic leadership*. 115-137

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education*, 5, 193– 211.

Nigerian Educational Research and Development Council, (2013). Senior secondary school mathematics curriculum for SS one to three, Abuja. Nigeria.

Oladele, B. K. and Adegoke, B. A. (2020). Using test theories models to assess senior secondary students' ability in constructed-response mathematics tests. *Journal of Education and Practice.*11: 46 -55 www.iiste.org.

Oladele B. K. (2021): Comparison of secondary school students' mathematics ability in mathematics constructed-response items under classical test and item response measurement theories in the Ibadan Metropolis, Nigeria. A PhD Thesis, Institute of Education, University Of Ibadan. ir.uiowa.edu.

Rizopoulos, D. (2006). ltm: An R Package for latent variable modelling and item response Theory Analyses. Journal of Statistical Software 17: 1-25.

Steven, J.P. (2012). Applied Multivariate Statistics for the Social Sciences, Fifth Edition. 664

Reckase, M. D. (1985). The difficulty of items that measures more than one ability/ *Applied ability/Applied Psychological Measurement* 9:401-412.

Stone, C. A., & Yeh, C.C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. Educational and Psychological Measurement, 66, 193–214.

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment*, 19, 35–57

Tate, R. 2003. A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159–203.

Watkins, D. S., (2006). Product eigenvalue problems *(.pdf), SIAM Review* 47:3-40. epubs.siam.org

Zhang. M., (2016). Exploring the dimensionality of scores for mixed-format tests. PhD\(Doctor of Philosophy) thesis, University of Iowa, https://doi.org/10.17077/etd.

Zopluoglu, Cengiz and Davenport Jr., Ernest C. (2017). A note on using eigenvalues in dimensionality assessment. *Practical Assessment, Research and Evaluation*, 22(7).

Beltz Verlag. van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. 2004. A comparative study of test data dimensionality assessment procedures undenonparametric IRT models. *Applied Psychological Measurement*, 28, 3–24