International Journal of Evaluation and Research in Education (IJERE)
Vol. 9, No. 2, June 2020, pp. 385~393
ISSN: 2252-8822, DOI: 10.11591/ijere.v9i2.20454

Dimensionality assessment of binary response test items: A non-parametric approach of bayesian item response theory measurement

Musa Adekunle Ayanwale[1], Flourish O. Isaac-Oloniyo[2], Funmilayo R. Abayomi[3]
[1]Education Foundations, Kampala International University, Uganda
[2]Institute of Education, University of Ibadan, Nigeria
[3]Babcock University High School, Nigeria

Article Info

Article history:

Received Dec 25, 2019
Revised Feb 2, 2020
Accepted Mar 16, 2020

Keywords:

Full information factor analysis
Item response theory
Multidimensional test
Test dimensionality
Unified mathematics achievement test
Within-item multidimensionality

ABSTRACT

This study investigated dimensionality of Binary Response Items through a non-parametric technique of Item Response Theory measurement framework. The study used causal comparative research type of non-experimental design. The sample consisted of 5,076 public senior secondary school examinees (SSS3) between the age of 14-16 years from 45 schools, which were drawn randomly from three senatorial districts of Osun State, Nigeria. Instrument used for this study was 2018 Osun State unified multiple-choice mathematics achievement test items with empirical reliability coefficient of 0.82. Data obtained were analysed using Non-linear factor analysis, Stout's Test of Essential Unidimensionality (STEU), Factor Analysis (FA), Full Information Factor Analysis (FIFA) and Bootstrap Modified Parallel Analysis Test (BMPAT). Results showed that both the BMPAT and STEU ascertained violation of unidimensionality assumption of the test items (the observed difference in the second eigenvalue of the observed data and that of second eigenvalue of the simulated data was statistically significant, $p = 0.0099$; Stout's test rejected the assumption of essential unidimensionality, $T = 10.6260$, $p<0.05$). Non-linear factor analysis and full information factor analysis revealed that four dimensions embedded in the test items and loadings of the items showed within-item multidimensionality respectively. The authors' concluded that modeling examinees' performance with unidimensional model when it was actually multidimensional in nature would affect performance of examinees adversely and could lead to blur judgment. Consequently, it is recommended that unidimensional scoring method of Osun State unified mathematics achievement test implicit in Classical Test Theory should be jettisoned and an appropriate scoring model (multidimensional) should be embraced.

Corresponding Author:

Musa Adekunle Ayanwale,
Education Foundations,
Kampala International University,
Box 20000, Ggaba Road, Kansanga, Kampala, Uganda.
Email: kunleayanwale@gmail.com

1. INTRODUCTION

Unified Examinations conducted for the transition of students from senior secondary school II to III was no doubt a laudable policy of the government of Osun State, conceptualised to enhance the performance of students at both internal and external examinations. This noble idea was geared towards assessing

the abilities of students through valid and reliable test items in different subjects including Mathematics. Mathematics is a brilliant vehicle for the advancement and enhancement of an individual's intellectual competence in logical thinking, spatial perception, analytical and abstract idea. It is the language for quantitative description of the world and it is an essential tool in many fields such as engineering, medicine, science and applied science. Mathematics as a subject has been a compulsory subject at the primary and secondary schools in Nigerian for many years. The central position given to it might be because of the roles it plays in career pursuits of students, and also a pre-requisite for admission to study many courses at the post-secondary level of education.

In the field of educational measurement all over the world, had in the last four decades experienced changes to satisfy expanding needs for valid interpretations of individual scores from educational tests. Measurement frameworks that are widely used for scoring of students' test performance is popularly called the Classical Theory (CTT) and Item Response Theory (IRT). IRT which take into cognisance the dimensionality (that is uni-dimensional or multi-dimensional) of the test data and model the performance of students with the number of evident dimensions. However, this measurement framework is distinct from Classical Test Theory (CTT) framework which assumed that scoring of students' performances can only be done within the confine of unidimensional assumption. That is, when a single trait is enough to explain the variation observed in the student's performance.

Test dimensionality can be described as the number of traits underlying a test that accounts for variation in students' test performance. The student's performance, in an extremely set of n-dimensional latent space are often described by a vector of ability scores as (θ_1, θ_2, θ_3…. θ_n). Item response models that presume many latent traits which accounts for students' test performance is known as multidimensional item response model, whereas item response models that assume a single latent ability is known as unidimensional models [1, 2]. This implies that the test items assess single area of construct or knowledge or ability. Adewale, Bejar, and Lynch [3-5] contend that a unidimensional test comprises of items that tap into only one dimension. Whenever a reported score for a test is single, there is an implicit presumption that the items share a typical primary construct or ability. This implies that the model includes a single θ for each and every examinee. Any other factor affecting examinee response to an item is regarded as random error or nuisance dimension distinct to that item and not shared by other items. The idea of unidimensionality therefore requires that all the items on a test or ability scale must measure a single latent trait of an individual. Violation of this would lead to modeling the students test performance with multidimensional models.

The nature of test data determines which of the available techniques should be used for assessing the dimensionality. When data are presumed to be continuous and normally distributed, dimensionality is determined using parallel analysis which is implemented in VistaParan and MPLUS, or Confirmatory Factor Analysis based on Pearson's correlation matrix [3, 6, 7] implemented in AMOS or LISREL respectively. When the data is ordinal, parallel analysis based on polychoric correlation (implemented in FACTOR; Vista-Paran) can be used. But when the data is dichotomously scored, nonlinear factor analysis implemented in Normal Ogive Harmonic Robust Moment (NOHARM), parallel analysis based on tetrachoric correlation matrix (implemented in Vista-Paran), Full Information Item Factor Analysis (implemented in EQSIRT, MIRT R package and TESTFACT), Bootstrap Modified Parallel Analysis Test (implemented in Itm (UnidimTest) R package) and Stout Essential Dimensionality Test (implemented in DIMTEST package) can be used [8-11]. In this study, Stout Test of Essential Dimensionality Analysis, Normal Ogive Harmonic Robust Moment, Bootstrap Modified Parallel Analysis Test and Full Information Item Factor Analysis were used. This would enable the researchers to cross-validate outputs generated from different techniques of assessing the dimensionality of a test data.

Stout Test of Essential Dimensionality Analysis (STEU) can be achieved by splitting an item pool into two separate clusters, then evaluating the distinctness of the responses in each cluster. The first of these clusters, called the assessment subtest (AT), is chosen such that the items contained within the partition are homogenous to one another, but as dimensionally distinct from the remaining items as possible. The second cluster is called partitioning subtest (PT), which consists of all items not used in assessment subtest (AT). It is used to cluster examinees based on their total partitioning subtest (PT) score. The separation of items into these two clusters can be done using a variety of approaches ranging from those stemming from the factor analytic tradition [12] to clustering algorithms employing conditional covariance-based assessments of dimensional distinctness [13]. These partitioning methods can be approached in an exploratory or confirmatory manner. Thus, this study used an exploratory partitioning approach of DIMTEST. The null and alternative hypotheses tested by DIMTEST are given by [14]. They are:

H_0: $AT \cup PT$ satisfies essential unidimensionality ($d = 1$)

H_i: $AT \cup PT$ fails to satisfy $d = 1$

The null hypothesis posits that the AT and PT partitions assess the same dominant underlying dimension, while the alternative hypothesis implies that the items in the AT partition are best represented by

a dimension that is distinct from that driving responses to the PT items. After, correction for bias, the test statistic, T, is assumed normally distributed. The null hypothesis for T is that the responses are unidimensional, thus failure to reject the null hypothesis signifies that the assumption of unidimensionality is justifiable. If otherwise, multidimensionality manifest [15]. Another method adopted in this study for the assessment of dimensionality analysis of test data is Normal Ogive Harmonic Robust Moment (NOHARM). This feat can be achieved through the following steps:

a. Data formatting (that is data should be saved in fixed ASCII (.dat)) to conform with NOHARM environment
b. Open the NOHARM environment
c. Click the button to specify the Input File name.
d. Click the button to specify the Output File name
e. If desired, click the button to specify additional options. Here the researcher can:

– adjust the criterion for convergence. By default, this value is 0.000001, meaning that the job will be considered to have converged when the largest absolute gradient in the minimisation procedure is less than this value. (Theoretically, all gradients should be zero, but this is rarely achieved numerically).
– select a value for the maximum number of function calls. By default, this value is 2000. If a job does not converge, it might require re-running with a larger number of function calls allowed.
– select to have the program print the sample product-moment matrix to a separate file. This file will be given a name identical to the output file name, but will be preceded by "PM_". For example, if the output file is Outfile.txt, this file will be named PM_Outfile.txt.
– select to have the program print the residual matrix printed to a separate file. This file will be given a name identical to the output file name, but will be preceded by "RES_" appended. For example, if the output file is Outfile.txt, this file will be named RES_Outfile.txt.

f. Click the Start the Job button

– The Start the Job button will become an abort the job button until the job has finished.
– The decreasing values of the loss function and the largest gradient will be displayed until the job has finished, at which point the final values will be shown. These numbers can be displayed in exponential format. When the job is completed,

g. Click the view output file button in order to see the output file.

However, in practice, the data would be calibrated under the hypothesis that one dimension fitted the test data. Thereafter, the data set would be calibrated under the hypothesis that two dimensions underlie the test and then the fitness of one and two dimensions would be compared. If two dimensions fit the data better than one dimension, the data would be further calibrated under the hypothesis that three dimensions fit the data set and the fitness of two and three dimensions would be compared. Also, if indices of three dimensions fit the test data better than two dimensions, the data would be further calibrated under the hypothesis that four dimensions fit the data set and the fitness of three and four dimensions would be compared. Thus, the same trend would be followed until the optimal dimension embedded in the model is ascertained. More importantly, in cases where there are many dimension models with an acceptable level of fit indices, the most parsimonious model is determined by calculating the percentage reduction in RSMR values of the models being compared. As remarked by [16] that model with highest dimension that still produces an approximately 10% greater than or equal decrease in the RMSR over the preceding model is the model that fit the data most. Consequently, the percentage reduction in RMS for n-dimension to m-dimension can be expressed as:

$$\frac{(RMS\ n-\text{dim}ension\ \text{mod}el\ -RMS\ m-\text{dim}ension\ \text{mod}el)}{RMS\ n-\text{dim}ension\ \text{mod}el}\times 100 \quad (1)$$

Another approach is called Bootstrap Modified Parallel Analysis Test (BMPAT). This function implements the procedure proposed by [17] for examining the latent dimensionality of dichotomously scored item responses. The statistic used for testing unidimensionality is the second eigenvalue of the tetrachoric correlations matrix of the dichotomous items. A Monte Carlo procedure is used to approximate the distribution of this statistic under the null hypothesis that the second eigenvalue of the observed data is not substantially larger than the second eigenvalue of data under the assumed IRT model. Thus, violation of the null hypothesis depicts that the test data are inherently multidimensional in nature.

Once there is manifest of multidimensionality in test data, the next point of call is to assess the number of traits embedded in the data. This information provides a group for the choice measurement model with which test ability can be estimated [18]. To assess the number of dimensions embedded the test, parallel analysis is often used [10], which was designed for carrying out Exploratory Factor Analysis (EFA) for data that are continuous. Dichotomous data are not continuous and cannot be presumed to follow

multivariate normality. Thus, nonlinear factor analysis was suggested by psychometricians as panacea [19]. A variant of nonlinear factor analysis is full-information factor analysis, which is said to improve upon various traditional factor analyses by jointly estimating multiple thresholds and factor loadings for each item [20, 21]. It was developed for binary item responses. In order to determine the number of dimensions embedded in test data by comparing the fit indices of the model as the number of dimensions are increasing. Once the optimal model is achieved, the nature of the dimensions embedded the data is investigated using factor loadings, and this would establish test items that loaded together on different factor. Tabachnik [22] remarked that a factor is regarded as viable, only if it has more than the three items loading with minimum benchmark of 0.32. Thus, item loading with value of 0.32 and above is considered adequate.

Moreover, the nature of multidimensionality evident in the test data needs to be examined. Adams [23] suggested that there are two possible multidimensional models: A between-item multidimensional model and a within-item multidimensional model. In the between-item multidimensional case, each of the items is associated with a single dimension with several subscales. For instance, performance in the reading and listening items is modeled by two dimensions. A test is regarded as within-item multidimensional model if items are associated with more than one dimension that is such item requires ability from more than one dimension. Perusal of literature such as [24-26] had shown that numerical test items such as quantitative reasoning, mental ability achievement test etc. were generally believe to measure more than one traits that is they are multidimensional in nature. However, test items used for Osun State unified examination are inherently multidimensional. Scoring of examinees seems to be done within the confine of Classical test theory which is presumed to be unidimensional. Oguoma [26] argued that modeling examinees' performance with unidimensional model when is actually multidimensional would affect the performance of the examinees adversely. The question now is whether the dimensionality of Osun State Unified Examination test items is assessed? Therefore, this study investigated dimensionality assessment of binary response item: A non-parametric approach of Bayesian item response theory measurement.

2. RESEARCH QUESTIONS

Three questions guided this study: (1) Are Osun State unified Mathematics test items essentially unidimensional? (2) How many number of dimensions accounted for the observed variation in examinee responses to Osun State unified Mathematics test items? (3) What is the nature of the dimensionality of Osun State unified Mathematics test items?

3. METHODOLOGY

The study used causal comparative research type of non-experimental design. The population used for the study comprised of senior secondary school three examinees' (SSS3) from 245 public co-educational schools in Osun State. The sample for the study consisted of 5,076 SSS3 examinees. Osun State was classified along the existing three senatorial districts. In each of senatorial district, fifteen (15) schools were drawn using simple random sampling technique, making a total of 45 schools. An intact class of SSS 3 examinees in all the selected schools participated in the study. Among the 5076 examinees sampled, 3299 (65.0%) were boys, while 1777 (35.0%) were girls. Their ages ranged between 14 and 16 years. Adopted instrument was used for this study. This was year 2018 binary response Mathematics Achievement Test items of Osun State unified examination with empirical reliability coefficient of 0.82. The test contained fifty (50) items with four response format options, which were scored binary as 0 and 1. Data obtained was analysed using Non-linear factor analysis, Stout's Test of Essential Unidimensionality (STEU), Factor Analysis (FA), Full Information Factor Analysis (FIFA) and Bootstrap Modified Parallel Analysis Test (BMPAT).

4. RESULTS

4.1. Research question 1: Are Osun state unified mathematics test items essentially unidimensional?

This question was answered using Bootstrap modified parallel analysis test (BMPAT) implemented in ltm (UnidimTest) of R package and Stout's test of essential unidimensionality implemented in DIMPACK. The two methods were used in order to establish comparability of results obtained. The result of BMPAT is presented in Table 1 and Figure 1 respectively. Cursory look at Table 1 and Figure 1 showed that the second eigenvalue of the observed data (2.5946) was substantially larger than the second eigenvalue of the simulated data (0.3795) under the assumed 3PL model of Item response theory (IRT). More so, the bootstrap modified parallel analysis showed that the difference observed was statistically significant ($p = 0.0099$). Therefore, the null hypothesis was rejected, hence, the 2018 Osun State unified Mathematics test items were not unidimensional. It means that multidimensional is evident in the test items.

Table 1. Bootstrap modified parallel analysis test of dimensionality

	Value	P-value
Second eigenvalue in observed data	2.5946	0.0099
Average of second eigenvalues in Monte Carlo samples	0.3795	

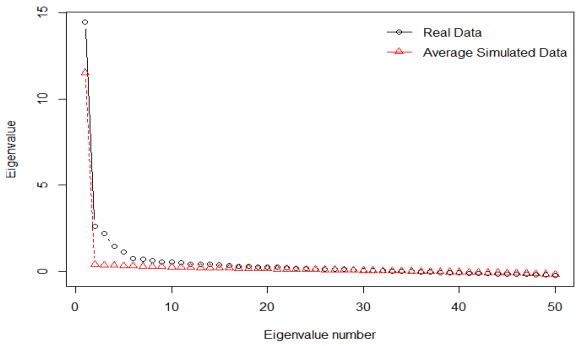

Figure 1. Parallel analysis of Osun state unified mathematics test items

Stout's test of essential unidimensionality was examined. This method provided a statistical test of the null hypothesis of essential unidimensionality. Stout [12] remarked that essential unidimensionality holds when the mean absolute value of the pairwise item covariances, conditional on Θ, is approximately zero (0). To achieve this feat, the items were divided into two subtests that were as dimensionally distinct as possible, the Partitioning Subtest (PT) and the Assessment Subtest (AT). The Assessment Subtest comprised of items that were potentially measured a secondary dimension. This approach empirically used sampled 30% of the examinees to search for clusters of items that might be influenced by a secondary dimension. In using 30% of the sampled examinees, items 6, 7, 11, 12, 16, 21, 24, 25, 26, 27, 28, 29, 30, 35, 38, 41, 42, 43 and 49 were empirically detected by the HCA/CCPROX clustering approach as the likely cluster to form a secondary dimension. The null and alternative hypotheses tested by DIMTEST are given by [14]. They are:
H_0: *AT* ∪ *PT* satisfies essential unidimensionality (d = 1)
H_i: *AT* ∪ *PT* fails to satisfy d = 1

The null hypothesis posits that the assessment subset (AT) and partitioning subset (PT) partitions assess the same dominant underlying dimension, while the alternative hypothesis implies that the items in the AT partition are best represented by a dimension that is distinct from that driving responses to the PT items. Table 2 presented DIMTEST statistic of essential unidimensionality. Table 2 revealed that Mathematics test items violated essential unidimensionality assumption with (T= 10.626, p = 0.000), and null hypothesis was rejected. This led to the conclusion that the AT items were dimensionally distinct from the remaining items in PT. Therefore, multidimensionality is manifest. More so, this result validated earlier finding obtained from BMPAT that Osun State unified Mathematics test items were implicit multidimensional in nature.

Table 2. Stout's test of essential unidimensionality test

TL	TGbar	T	P-value
24.6265	13.9475	10.6260	0.0000

4.2. Research question 2: How many number of factors accounted for the observed variation in examinee responses to Osun state unified mathematics test items?

To answer this question, examinee responses were subjected to full information factor analysis implemented in MIRT package of R programming language and non-linear factor analysis implicit in Normal Ogive Harmonic Analysis Robust Method (NOHARM) package [27] for cross validation of results obtained. Examinee responses were subjected to non-linear factor analysis implicit in Normal Ogive Harmonic Analysis Robust Method. In NOHARM analysis, there are two possible indices for the assessment of test dimensionality. The first index is the Root Mean Square Residual (RMSR) and the second is Tanaka's Goodness-of-Fit Index (GFI). McDonald [19] submitted that an hypothesised dimensional model is acceptable, if its GFI value is 0.90; and the model is regarded good if GFI value is ≥ 0.95, and perfect when GFI value = 1. RMSR Criterion was used as cut-off through which the general fitness of a model is judged.

The RMSR is the square root of the average squared difference between the observed and predicted covariance. Thus, as remarked by McDonald [28] that small value of RMSR shows good fit. Because small value is relative, the RMSR value is compared to a criterion which is four times the reciprocal of the square root of the sample size which is the "typical" standard error of the residuals. RMSR value is regarded as good when it is lesser than the criterion. Consequently, an hypothesised model is considered good, if the GFI and RMSR are good. Table 3 presents the results from NOHARM analysis.

Table 3. NOHARM dimensionality analysis of unified mathematics test

Dimension	GFI	RMSR Criterion	RMSR	Difference in RMSR	Reduction RMSR	% Reduction
1	0.857	0.08081	0.00997			
2	0.972	0.08081	0.00814	0.00183	0.18355	18.3
3	0.981	0.08081	0.00674	0.00140	0.01719	17.2
4	0.984	0.08081	0.00604	0.00070	0.10385	10.4
5	0.986	0.08081	0.00557	0.00047	0.07781	7.8

Table 3 revealed that the entire hypothesised dimension model (that is one-dimension model, two-dimension model up to five-dimension model) fitted the data. Meanwhile, when GFI of the models used in assessing the fitness of the data, one-dimension model did not fit the data (GFI < 0.90). This is an indication that unidimensionality assumption had been violated by the test data. The results further showed that from one-dimension to two-dimension, RMSR value decreased by 18.3%. This as suggested by Tate [16] criteria, is a significant benchmark for reduction. Also, from two-dimension model to three-dimension model, RMSR value decreased by 17.2%. Similarly, from three-dimension model to four-dimension model, RMSR value decreased by 10.4%. However, the reduction of RMSR from four-dimension to five-dimensions was 7.8%. This value was less than the criterion of 10% set by Tate. Therefore, the highest dimensional model that still produced an approximately 10% or greater percentage reduction in the RMSR over the preceding model was the four-dimensional model. The study concluded that four-dimensional model was the most parsimonious model that fitted the data better.

More importantly, examinee responses were also subjected to full information item factor analysis implemented in MIRT package of R programming language for cross validation of results. Seven factors model were hypothesised and compared using (Akaike Information Criterion (AIC), Akaike Information Criterion corrected (AICc), Bayesian Information Criterion (BIC), Sample size adjusted Bayesian Information Criterion (SABIC) and Likelihood ratio test (LogLik) in order to establish the best fit model that provided the information for the exact number of factors embedded in the test data. The results are presented in Table 4.

Table 4. Dimensionality analysis of unified mathematics achievement test

Dimension	AIC	AICc	SABIC	BIC	LogLik	X^2	df	P
comparing the fitness of two and three dimension models								
2	264191.6	264200.6	264691.4	265164.9	-131946.8	1790.659	48	0.000
3	262596.9	262622.3	263425.5	264210.4	-131051.5			
comparing the fitness of three and four dimension models								
3	262596.9	262622.3	263425.5	264210.4	-131051.5	628.743	47	0.000
4	258246.8	258297.0	259400.8	260493.9	-128779.4			
comparing the fitness of four and five dimension models								
4	258246.8	258297.0	259400.8	260493.9	-128779.4	433.824	46	0.000
5	257710.1	258775.2	259602.3	260257.7	-128465.0			

Table 4 showed the analysis of number of dimensions embedded in Osun State unified Mathematics Achievement Test. The fitness of hypothesised models was assessed using the information criteria with the most parsimonious model. First, when the fitness of hypothesised two and three dimension models were assessed, the result indicated that the three-dimension model had AIC, AICc, BIC and SABIC values less than that of two-dimension model. Thus, three-dimension model fitted the data better with statistical significant of likelihood ratio (χ^2 (48) = 1790.659, $p < 0.05$). Also, when the fitness of three-dimension model and four-dimension model were evaluated, four-dimension model had AIC, AICc, BIC and SABIC values less than that of three-dimension model with statistical significant of likelihood ratio (χ^2 (47) = 628.743, $p < 0.05$).

Furthermore, the result showed that when the fitness of four-dimension model was compared to five-dimension model, the trend of information criteria was not clear. Meanwhile, the likelihood ratio of five-

dimension model was statistically significance with (χ^2 (46) = 433.824, p < 0.05) but, less values of information criteria (AIC, AICc, BIC and SABIC) of four-dimension model fitted the data accurately. Consequently, it can be inferred from obtained results that four-dimension model fitted the data better than when compared with other hypothesised dimension models. Although, five-dimension model showed to fit the data better but there were some conflicting indices from the result that made the fitness not better. Therefore, parsimonious model of four-dimension fitted the Osun State unified Mathematics Achievement Test. The implication of this is that four different traits accounted for the observed variation in examinees performance in the test. This result also validated earlier finding obtained from NOHARM that Osun State unified Mathematics test items were embedded with four dimensions.

4.3. Research question 3: What is the nature of dimensionality of Osun state unified mathematics test items?

Here, the parsimonious model from NOHARM and full information item factor analysis (FIFA) respectively (four-dimension model) was used to determine the nature of dimensionality of the test. The resulting factor loading from the factor analysis conducted is presented in Table 5

Table 5. Nature of dimensionality osun state unified mathematics achievement test

Item	Factor 1	Factor 2	Factor 3	Factor 4
Item 1	0.169	0.177	0.117	0.598
Item 2	0.280	0.135	0.200	0.312
Item 3	0.772	-0.280	0.920	-0.536
Item 4	0.405	0.181	0.166	0.185
Item 5	0.355	0.211	0.209	0.389
Item 6	-0.195	0.550	0.563	0.104
Item 7	0.256	0.310	0.175	1.278
Item 8	0.319	0.168	0.128	0.112
Item 9	0.370	0.190	0.034	0.178
Item 10	0.315	0.612	0.168	-0.104
Item 11	0.276	0.367	-0.188	0.137
Item 12	0.570	0.544	0.125	1.114
Item 13	0.284	0.317	0.130	0.133
Item 14	0.462	0.124	0.120	0.108
Item 15	0.434	0.107	0.118	-0.114
Item 16	0.222	0.471	0.783	0.157
Item 17	0.364	0.139	0.286	0.147
Item 18	0.484	0.083	0.126	0.127
Item 19	0.462	0.243	0.125	0.187
Item 20	0.320	-0.254	-0.126	-0.105
Item 21	0.903	0.491	-0.188	0.402
Item 22	0.215	0.304	0.270	-0.282
Item 23	0.367	0.114	0.138	0.186
Item 24	0.299	0.309	0.209	0.128
Item 25	0.307	0.247	0.259	0.024
Item 26	0.152	0.454	0.104	0.190
Item 27	0.304	0.272	0.243	0.138
Item 28	0.256	0.306	0.255	0.021
Item 29	0.290	0.229	0.196	0.120
Item 30	0.243	0.318	0.142	0.348
Item 31	1.457	-0.179	0.963	-0.470
Item 32	0.498	0.052	0.230	0.236
Item 33	0.392	0.169	0.271	0.151
Item 34	-0.207	0.267	0.555	-0.443
Item 35	0.267	0.246	0.232	0.162
Item 36	0.347	0.134	0.286	0.142
Item 37	0.399	0.027	0.216	0.302
Item 38	0.208	0.422	0.215	0.307
Item 39	0.252	0.199	0.249	-0.809
Item 40	0.353	1.318	-0.024	-0.325
Item 41	0.306	0.212	0.263	0.195
Item 42	0.169	0.248	0.355	0.120
Item 43	0.022	0.435	0.346	0.136
Item 44	0.287	0.276	0.334	0.246
Item 45	0.209	0.319	0.490	0.712
Item 46	0.115	0.153	0.435	-0.013
Item 47	0.412	-0.131	0.401	0.289
Item 48	0.161	0.266	0.461	0.008
Item 49	0.534	0.374	0.459	0.105
Item 50	0.225	0.186	0.360	0.231

Table 5 showed the factor loading of unified Mathematics multiple-choice items after rotation method. As suggested by Tabachnik [22], substantial loadings were the highlighted red colour loadings with values greater than or equal to 0.32. Consequently, multidimensionality is evident, if an item has more than one substantial loading. It can be inferred from Table 5 that four factors (F1, F2, F3 and F4) out of five factors hypothesised had more than three items loading with values greater than or equal to 0.32, which is the benchmark for considering the viability of a factor. Cursory look at the loadings depicted that items 4, 9, 14, 15, 17, 18, 19, 23, 32, 33, 36 and 37 loaded on factor 1, items 10, 11, 26 and 38 loaded on factor 2, items 34, 42, 44, 46, 48 and 50 loaded on factor 3 while items 1 and 7 loaded on factor 4 respectively. In the same vein, items 3, 31, 47 and 49 cross-loaded on more than one factor (that is factor 1 and factor 3), items 5, 12 and 21 cross-loaded on factor 1 and factor 4, items 6, 16, 43 and 49 cross-loaded on factor 2 and factor 3 and items 12 and 21 cross-loaded on factor 2 and factor 4 respectively. More so, factor 1 and factor 2 had items 12, 21, 40 and 49 cross-loaded on them, item 45 loaded on factor 3 and factor 4, item 49 loaded on factor 1, factor 2 and factor 3 while factor 1, factor 2 and factor 4 had items 12 and 21 cross-loaded on them. More importantly, it was evident from these results that the nature of dimensionality of the Mathematics Achievement Test was multidimensional and the observed cross-loading was within-item multidimensionality.

5. DISCUSSION

Assessment of dimensionality of test item is the first point of call for any test developers to accomplish. Based on this, the results of this study revealed violation of dimensionality assumption of Osun State unified Mathematics Achievement Test items, otherwise multidimensionality was evident and the nature of dimensionality was within-item multidimensionality respectively. Findings from this study laid credence to the work of [26], who carried out dimensionality assumption test on 2014 Mathematics achievement items of West African Senior Secondary Certificate Examination (WASSCE). They concluded that the test items of WASSCE mathematics were inherently multidimensional in nature. Furthermore, the results of this study corroborated findings of [25] that fifty (50) items of 2013 WASSCE and sixty (60) items of National Examinations Council (NECO) Geography respectively violated assumption of unidimensionality and that there were more than one dimension that accounted for the variation observed in examinees to the geography test items. However, this result disagreed with the findings of [29] who concluded that Chemistry achievement test in the Kano State qualifying examination fulfilled the assumption of unidimensionality. The deviation observed between their study and this current one might be due to different statistical approach adopted. In their study, factor analysis was conducted to assess the dimensionality of the test data using Statistical Package for Social Sciences, version 20 (SPSS), which was designed for continuous data for factor analysis. Gadermann [30] suggested that using Statistical Package for Social Science (SPSS) for binary response items factor analysis, would substantially led to overestimation of the eigenvalues. In this study, full information item factor analysis which was designed for binary response items implemented in R programming language was used.

6. CONCLUSION

The study concluded that modeling examinees' performance with unidimensional model when it was actually multidimensional in nature would affect the performance of the examinees adversely and could lead to blur judgment. Therefore, it is recommended that unidimensional scoring method of Osun State unified mathematics achievement test should be jettisoned and an appropriate scoring model (such as multidimensional models) should be incorporated.

ACKNOWLEDGEMENTS

The authors express their profound gratitude to the director of schools in the ministry of education, State of Osun, Nigeria for granted our request and directed principals in all the selected schools to give us maximum cooperation during data collection for the study.

REFERENCES

[1] H. Hambleton and R.K. Swaminathan, *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff, 1985.
[2] A. L. Comrey and H. B. Lee, *A first course in factor analysis*. Hillsdale, NJ: Erlbaum, 1992.
[3] M. A. Adewale, *et al.*, *A training manual on item response theory*, 1st ed. Ibadan: Institute of Education, University of Ibadan in Collaboration with National Examinations Council, Minna, Niger State, 2017.

[4] I. Bejar, "A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates," *J. Educ. Meas*., vol. 17, no. 4, pp. 283–296, 1980.
[5] B. Lynch, "The factor structure of the Winter 1985 ESLPE," University of California at Los Angeles, 1985.
[6] J. Stevens, *Applied Multivariate Statistics for the Social Sciences*, 3rd Edition. Hillsdalc: Lawrence Erlbaum, 1996.
[7] R. B. Kline, "Software programs for structural equation modeling: Amos, EQS and LISREL," *J. Psycho-educational Assess*., vol. 16, pp. 343–364, 1998.
[8] W. Stout, "A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation," *Psychometrika*, vol. 55, no. 2, pp. 293–325, 1990.
[9] R. J. De Ayala, *The Theory and Practice of Item Response Theory*, 1st ed. New York, NY: The Guilford Press, 2009.
[10] M. D. Reckase, *Multidimensional Item ResponseTheory*. New York: Springer Verlag, 2009.
[11] P. Finch and W.H. Monahan, "A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment," *Appl. Meas. Educ*., vol. 21, pp. 119-140, 2008.
[12] W. F. Stout, "A nonparametric approach for assessing latent trait unidimensionality," *Psychometrika*, vol. 23, pp. 589–617, 1987.
[13] W. Zhang and J. Stout, "The theoretical DETECT index of dimensionality and its application to approximate simple structure," *Psychometrika*, vol. 64, no. 2, pp. 213–249, 1999.
[14] J. Stout, W. Habing, B. Douglas, J. Kim, H. R. Roussos, and L. Zhang, "Conditional covariance-based nonparametric multidimensionality assessment," *Appl. Psychol. Meas*., vol. 20, no. 4, pp. 331–354, 1996.
[15] C. DeMars, *Item Response Theory, Understanding statistics measurement*. NY: Oxford University Press, 2010.
[16] R. Tate, "A comparison of selected empirical methods for assessing the structure of Responses to test items," *Appl. Psychol. Meas*., vol. 27, no. 3, pp. 159–202, 2003.
[17] R. Drasgow and F. Lissak, "Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses," *J. Appl. Psychol*., vol. 68, pp. 363–373, 1983.
[18] L. Ying, "Applying Multidimensional Item Response Theory Models in Validating Test Dimensionality: An Example of K–12 Large-scale Science Assessment," *J. Appl. Test. Technol*., vol. 13, no. 2, pp. 220–239, 2012.
[19] R. P. McDonald, *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum, 1999.
[20] E. Bock, R.D. Gibbons, and R. Muraki, "Full-Information Item Factor Analysis," *Appl. Psychol. Meas*., vol. 43, pp. 293–322, 1988.
[21] A. Yu, C.H. Osborn Popp, S. DiGangi, and S. Jannasch-Pennell, "Assessing unidimensionality: A comparison of Rasch Modeling, Parallel Analysis, and TETRAD," *Pract. Assessment, Res. Eval*., vol. 12, pp. 14–23, 2007.
[22] L. S. Tabachnik and B.G. Fidell, *Using Multivariate Statistics*. Boston: Pearson, 2007.
[23] W. Adams, R.J. Wilson, and M. Wang, "The multidimensional random coefficients Multinomial logit model," *Appl. Psychol. Meas*., vol. 21, no. 2, pp. 1–23, 2009.
[24] B. Oyeyemi, "Modeling Undergraduates' Entry Requirements and Academic Performance in the Faculty of Science, University of Ibadan," *Ibadan*, October 17, 2019.
[25] M. A. Okwilagwe and E.A. Ogunrinde, "Assessment of Unidimensionality and Local Independence of WAEC and NECO 2013 Geography Achievement Tests.," *African J. Theory Pract. Educ. Assess*., vol. 5, pp. 31-44, 2017.
[26] R. O. Oguoma, C.C. Metibemu, and M.A. Okoye, "An Assessment of the dimensionality of 2014 West African secondary school examination mathematics objective test scores in Imo State, Nigeria," *African J. Theory Pract. Educ. Assessment*, vol. 4, pp. 18-33, 2016.
[27] R. P. Fraser and C. McDonald, "NOHARM: A windows program for fitting both Unidimensional and multidimensional normal Ogive models of latent trait theory {Computer Program}," *Niagara College, Welland*, 2003.
[28] R. P. McDonald, *Normal-Ogive multidimensional model*. New York: Springer, 1997.
[29] A. Bichi and A.A. Bichi, "Analysis of Dichotomous Scored Science Achievement Test Items using Item Response Theory Framework," in *Public Examining in Sub-saharan Africa: Issues, Challenges and Prospects*, O. E. E. Adegoke, B.A. Popoola, and O. Babatunde, Ed. Abuja: Joint Admission and Matriculation Board (JAMB), pp. 51–66, 2016.
[30] B. D. Gadermann, A.M. Guhn, and M. Zumbo, "Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide," *Pract. Assessment, Res. Eval*., vol. 17, pp. 1–13, 2012.