AN ASSESSMENT OF THE DIMENSIONALITY OF 2014 WEST AFRICAN SECONDARY SCHOOL CERTIFICATE EXAMINATION MATHEMATICS OBJECTIVE TEST SCORES IN IMO STATE, NIGERIA

Chinyere C. OGUOMA[a], Michael Akinsola METIBEMU[b], Romy OKOYE[c]
[a]Department of Psychology, Alvan Ikoku Federal College of Education, Owerri, Nigeria
chinyerechyzy@yahoo.com +2348036745446

[b] corresponding author, Institute of Education, University of Ibadan
metimike@gmail.com +2347058539888

[c] Department of Educational Foundations, Nnamdi Azikiwe University, Awka, Nigeria
romyokoye@yahoo.com +2348037305856

Abstract

Mathematics tests used in the assessment of students' proficiency in mathematics are inherently multidimensional. However, the procedure being adopted by public examining bodies in the scoring of examinees' performance in the subject is based on Classical Test Theory (CTT), a measurement model that is limited to assessing tests that are unidimensional. This may be one of the reasons why students are consistently performing poorly in Mathematics, as research has shown that assessing multidimensional tests using unidimensional test models compromises test performance. Hence, research in Mathematics education should shift focus from the dominant students, teachers, school related factors and Mathematics achievement towards scoring procedures being adopted by public examining bodies. Therefore, this study assessed the appropriateness of scoring 2014 WASSCE Mathematics multiple choice test using CTT. To achieve this, the dimensionality of the test was assessed. The study adopted causal comparative design. All the year three senior school students (SSS3) of the 274 public Senior Secondary Schools in Imo State that registered candidates for the May/June 2015 Senior School Certificate Examination (SSCE) formed the population for the study. The sample for the study comprised 1142 (SS3) students of 30 schools which were randomly selected from the 274 public senior secondary schools in the State. Data was analyzed using Stout's Test of Essential Unidimensionality, Bootstrap Modified Parallel Analysis test (BMPAT) and full information factor analysis. Results showed that the test and its items violate unidimensionality assumption (Stout's Test rejected the assumption that the test was unidimensional, $T = 7.3799$, $p < 0.001$, one tailed; BPMAT showed that the second eigenvalues of the real was significantly greater than that of the simulated data set, $p=0.0099$). Furthermore, full information factor analysis showed that five dimensions underlie the test; the test items revealed within-item multidimensionality. The findings suggest that 2014 WASSCE Mathematics test was multidimensional. It is therefore our conclusion that CTT scoring was not suitable for the measurement of students' performance in 2014 WASSCE Mathematics test. Hence, it is our recommendation that the use of CTT in the estimation of students' test performance in SSCE Mathematics tests be stopped.

Keywords: Classical test theory, Item response theory, test dimensionality, multidimensional test, within-item multidimensionality

Introduction

Mathematics, a body of knowledge that deals with quantity, structure, size and space, is the basic vehicle upon which science and technological advancement rides. Mathematics as a school

subject affects all aspects of human life at different degrees. For instance, mathematics is relevant in economics, political, geographical, scientific and technological aspects of man. Other areas where the use of numbers is predominant include; statistics, account, arithmetic, engineering and so on. In fact, the earliest civilization of mankind came through mathematical manipulations involving the use of numbers.

In spite of the importance of mathematics, the performance of students in the subject at the Secondary School Certificate level in Nigeria has not been encouraging (see Table1)

Table 1: Students' enrolment and performance in WASSCE May/June Mathematics (2003 – 2014)

Year	Total Enrolled for Examination	A1-C6 Credit Passes	% Credit Passes	D7-E8 Passes	% Passes	F9 Failure	% Failure
2003	1,038,809	341928	32.92	331,348	31.90	229,878	22.13
2004	1,035,266	287,484	34.52	245,071	29.43	300,134	36.05
2005	1,054,853	402,982	38.20	276,000	25.36	363,055	34.41
2006	1,181,515	482,123	41.72	366,801	31.55	292,560	25.13
2007	1,249,028	583,921	46.75	333,740	26.72	302,764	24.24
2008	1,292,890	726,398	57.28	302,266	23.83	218,618	17.23
2009	1,373,009	634,382	47.04	344,635	25.56	315,738	23.41
2010	1,306,535	548,065	41.95	363,920	26.85	355,382	27.20
2011	1,508,965	608,866	40.40	474,664	31.50	421,412	27.90
2012	1,550,224	723,024	46.64	445,224	28.72	380,425	24.54
2013	1,399,178	618,996	44.24	371,202	26.53	406,181	29.03
2014	1,547,140	621,950	40.20	427,342	30.53	451,301	29.17

As shown in Table 1, in 2003, the level of performance of students in Mathematics was below average; less than 50% of the students who sat for mathematics passed at credit level. The same trend was observed in 2004, 2005, 2006, and 2007. In 2008, the percentage pass at credit level rose above 50%. However, it was not sustained. The consistent poor performance observed between 2003 and 2007 was also observed in 2009, 2010 and 2012, 2013 and 2014. A cursory look at the table revealed that on the average about 58% of students who sat for Mathematics in WASSCE between 2003 and 2014 could not obtain the minimum credit pass. This implies that about 58% of students who left Secondary School between 2003 and 2014 were not qualified for admission into tertiary institutions in Nigeria.

Many studies conducted in Nigeria have tried to isolate factors responsible for this observed poor performance. The identified factors include amongst others motivational orientation, self-esteem/self-efficacy, emotional problems, study habits, teacher consultation, poor interpersonal relationships amongst students (Aremu & Sokan, 2003), students poor attitude towards Mathematics (Bolaji, 2005), and poor teaching method adopted by teachers (National Mathematics Centre, NMC, 2009). To reduce the failure rate, many interventions were suggested. Prominent among the interventions is the NMC's Mathematics Improvement Programme (MIP) aimed at creating a new teaching methodology to enhance students' performance in Mathematics. Despite the intervention, the observed poor performance persists. This is evident in the consistent poor performance of students in WASSCE Mathematics after the introduction of the intervention in 2009.

An aspect which research in Mathematics education has not focused on much is the appropriateness of the assessment framework used in scoring students' performance in Mathematics at the SSCE level. Mathematics tests used in the assessment of students' mathematical proficiency inherently measure more than one trait. However, the scoring procedure, CTT, being adopted by public examining bodies in the scoring of examinees' performances in Mathematics test do so under the unidimensionality assumption. That is, the scoring procedure assumes that variations observed in examinees' test performance is accounted for by only one trait. This may be one of the reasons why students are performing poorly in Mathematics at the SSCE level, as studies (Reckase, 1985; Ansley& Forsyth, 1985; Reckase, Carlson, Ackerman & Spray, 1986) in other climes have shown that multidimensional tests scored with unidimensional scoring model results in compromised students' test-scores. It is as a result of this that research in Mathematics education in Nigeria, should shift focus towards the appropriateness of the assessment procedure adopted by public examining bodies in the scoring of students' performance in Mathematics.

In the assessment of students' performance in Mathematics at the Secondary School Certificate Examination (SSCE) level, two sets of achievement tests are used. These include: Multiple choice and Essay Mathematics tests. In the Essay test students are free to choose a required number of questions to answer from among the questions contained in the tests' booklet. This freedom of choice enables students to attempt questions they find appealing. Thus, making students' test scores comparison a very difficult task. In contrast, examinees are required to answer all questions contained in the multiple choice test. Thereby providing a level ground for examinees' test scores comparison. Multiple choice test assesses students' proficiency without necessarily elongating testing time. More importantly, it is highly objective when it comes to scoring of examinees responses. In this study, therefore, the Mathematics multiple choice test is emphasized.

In educational testing, another framework used in scoring students tests performance, be it multiple choice test or essay test is the Item Response Theory (IRT). This measurement framework unlike CTT assesses the dimensionality of a test and models test performance based on the dimensionality of the test. This may be one of the reasons why testing in the developed nations is based on IRT framework.

Test dimensionality refers to the number of trait underlying a test that accounts for variation in examinees test performance. A test is termed unidimensional if there is only one trait (dimension) accounting for variation in examinees test performance (Yu, Popp, DiGangi & Jannasch-Pennel, 2007). When there are two or more constructs accounting for variation in examinees test performance, the test is termed multidimensional. According to Tate (2003), the

assessment of the number of dimension resulting from the interaction of examinees with items in a test should be an important part of the development, evaluation, and maintenance of large-scale tests. This is because the assessment provides empirical support for the content and cognitive process aspects of test validity (American Educational Research Association; American Psychological Association & National Council on Measurement in Education, 1991).

Furthermore, the assessment helps to uncover the possible violations of the assumption of unidimensionality that is implicit in the assumption of homogenous items in Classical Test Theory (McDonald, 1999) and explicit in the version of IRT (unidimensionality IRT) currently in use in Nigeria (Metibemu, 2016; Adegoke 2013; Ojerinde 2013).This suggests that a test having more than one dimension cannot be accurately assessed using the CTT and unidimensional IRT. According to Reckase (2009), when a test measures more than one dimensions, multidimensional IRT is the appropriate statistical tool for the assessment of person and item statistics in the test. Therefore, for precise measurement, test dimensionality assessment is expedient.

In the assessment of test dimensionality, the first step usually adopted by test developers and evaluators is to assess the tenability of the unidimensionality assumption explicit in CTT and implicit in the unidimensional IRT (UIRT). In the assessment of test dimensionality, several methods are applicable. Prominent among these methods is the Stout's test of essential unidimensionality (Stout, 1987; Demars, 2003; De Ayala, 2009; Ackerman, Gieri & Walker, 2003). The method has been consistently used as the standard for validating newer parametric methods. For example, Finch and Monahan (2008), validated the effectiveness of bootstrap modified parallel analysistest (BMPAT) in assessing test dimensionality using the Stout's procedure. The result showed that the BMPAT was equally effective in the assessment of the undimensionality assumption of dichotomously scored test data. In the present study, the two methods, Stout's test of essential unidimensionality and BMPAT were used. The methods were used for the purpose of cross validation of results.

Essential dimensionality is based on the assumption that, there is one and only dominant latent ability influencing examinees' test performance. Essential dimensionality holds when the mean absolute value of the pairwise item covariance conditional on the underlying trait is approximately zero (Stout, 1987). This test is implemented in DIMTEST 2.0 software (Stout, 2005). To perform this test, the test items are divided into two subtests that are distinct as possible, the partitioning subtest (PT) and the assessment subtest (AT). The AT consists of items that measure the primary dimension and PT consists of the remaining items after the AT has been removed (this remaining items is assumed to measure the secondary dimension) measuring the secondary dimension This division can be done empirically using the inbuilt cluster algorithm of the software or done manually using the analysis of the test content. After correction for bias, the test statistic, T, is assumed normally distributed. The null hypothesis for T is that the responses are unidimensional (i.e., the average covariance within groups is zero), so failure to reject the null hypothesis indicates that the assumption of unidimensionality is tenable. If otherwise, multidimensionality is evident (Demars, 2010).

A second approach for assessing the assumption of unidimensionality in this study, the BMPAT, is an extension of modified parallel analysis (Budescu, Cohen, & Ben-Simon, 1997), which itself is based on Horn's (1965), parallel analysis (PA) method for determining the number of factors underlying a test data in factor analysis modeling.Implementation of the BMPAT requires a number of mathematical steps. These steps according to Finch and French (2015), are as follow:

Step 1. Estimate a unidimensional IRT model for the data

Step 2. Conduct an item factor analysis as described above, and save the eigenvalues.
Step 3. Simulate θ for N examinees based on the θ distribution for the actual data.
Step 4. Simulate unidimensional item responses using the θ values from step 3 and the item parameters from step 1.
Step 5. Conduct item EFA on the data simulated in step 4, and save the eigenvalues.
Step 6. Repeat steps 3–5 many (e.g. 1000) times.
Step 7. Compare the eigenvalues for each factor from the original data (step 2) with the distributions of eigenvalues for each factor created in steps 3-5. If the observed eigenvalue for a given factor is greater than or equal to the 95^{th} percentile of the simulated eigenvalue distribution, then conclude that the observed factor is present in the data.
Step 8. Repeat step 7 until the observed eigenvalue does not exceed the 95th percentile of the simulated eigenvalue for a given factor (pg, 237).
According to Finch and French (2015), a data is considered multidimensional, when on comparison of the second eigenvalues of the real and simulated data sets and the real eigenvalue is significantly greater than the 95^{th} percentile of the generated eigenvalue. If otherwise, then it is concluded that data are unidimensional.
If multidimensionality is evident in a test data, a second step in dimensionality assessment is to determine the number and nature of the dimensions present in the data. This information provides a group for the choice measurement model with which test performance can be estimated. To assess the number of dimensions underlying the test, parallel analysis is often recommended (Reckase, 2009; finch & French, 2015). However, PA being an EFA based was developed for data that are continuous in nature (and multivariate normal) (Finch & French, 2015). Clearly item response data are not continuous, and cannot be assumed to be multivariate normal. Indeed, researchers have found that using continuous data EFA with dichotomous item response data will result in the identification of spurious factors, leading to incorrect conclusions regarding data dimensionality (Hattie, 1985). Therefore, alternative approaches for fitting EFA models to the dichotomous item data have been proposed. For example, De Ayala (2009) and Finch and French (2015) suggested the fitting the data to multidimensional item response theory (MIRT) model using full information maximum likelihood expectation maximization (EM) algorithm of Bock and Aitkin (1981) or Metropolis–Hastings approach outlined by Cai (2010). The number of dimensions underlying the test data is obtained by comparing the fit indices of the model as the number of dimensions is increased.To achieve this feat several measures apply. Chief among the measures include Chi-square difference test and use of information indices (Finch & French, 2015). However, the information indices measures have been found to have some advantages over the chi-square difference test. In that the models do not need to be nested for comparisons to be conducted.
Information indices are simply measures of variance not explained by a model, with an added penalty for model complexity. Among the most popular of these indices are the -2loglikelihood (Kline, 2005), **Akaike information criterion** (AIC; Akaike, 1973), the **Bayesian information criterion** (BIC; Schwarz, 1978), and the **sample-size-adjusted BIC** (SBIC; Enders &Tofihi, 2008). Each of these statistics is based upon the model chi-square, and is interpreted such that the model with the lower value exhibits a better fit to the data.In addition, the chi-square and likelihood ratio goodness of fit tests the null hypothesis that two nested models provide the same fit to a set of data.A statistically significant likelihood indicates a difference in the models under examination.

Once the optimal model has been obtained, the nature of the dimensions underlying the data is examined using factor loadings; to ascertain which items appear to group together. According to Tabachnick and Fidel (2013) a dimension is considered substantial if it has three or more items having factor loading greater than or equal to 0.32. This thus implies that lading is considered substantial if its value is 0.32 and above. Another consideration is the issue of types of multidimensionality. To assist in the discussion of different types of multidimensional models and tests, Wang (1995),introduced the notions of *within-item* and *between-item multidimensionality*. A test is regarded as multidimensional between-item if it is made up of several unidimensional subscales. A test is considered multidimensional within-item if any of the items relates to more than one latent dimension.In psychometrics when items loaded on more than factor substantially, it is concluded that the items require abilities from more than one dimension.Such tests are called within-item multidimensional test (Adams & Wu, 2010).

In Nigeria, literature search showed only one study (Awopeju&Afolabi, 2016) that has assessed the dimensionality of SSCE Mathematics test. In the study, it was found that the Mathematics test was unidimensional. In the study, Awopeju and Afolabi assessed the unidimensionality of 2011 National Examinations Council Mathematics test, a dichotomous test, with the analysis of the eigenvalues obtained from the factor analysis of the test data using factor analysis module of SPSS, a factor analysis developed for continuous data.

Statement of the problem

The trend in performance of Nigerian students in the West African Secondary School Certificate Examination (WASSCE) Mathematics has not been encouraging. To improve on the performance trends, many empirical studies conducted in Nigeria recommended the use specialized teaching methodology among others. However, it appears that the recommended interventions were not very effective. This is because the observed level of performance persists even after the interventions. An aspect that is yet to receive full research attention in Mathematics education in Nigeria is the extent to which assessment practices can affect students' performance. No doubt, the examinees' performance in a test, which contains items that measure more than one latent trait or factors, will be adversely affected if scored using measurement framework that does not have the capability of modeling tests with more than one trait. In Nigeria, measurements of students' achievement at the SSCE level have always been based on CTT, a theory which assumes that tests measure only one latent trait. Another measurement framework often used in the developed nations for the measurement students' test performance is the Item response theory. This theory models test performance using the trait underlying the test. Therefore, this study assessed the dimensionality of the WASSCE 2014 Mathematics test.

Hypothesis

2014 WASSCE Mathematics multiple choice test is essentially unidimensional.

Research questions

How many traits underlying 2014 WASSCE Mathematics test account for variation in examinees responses?

What is the nature of 2014 WASSCE Mathematics test dimensionality?

Methodology

The study adopted causal comparative design. The population consisted of all the year three senior school students (SSS3) of the 274 public Senior Secondary Schools in Imo State that registered candidates for the May/June 2015 Senior School Certificate Examination (SSCE). The sample for the study comprised 1142 (SS3) students. In the selection of the sample, simple

random sampling technique was used to select 10 schools from each of the three educational zones into which the secondary schools in the state were divided. Thus, 30 senior secondary schools were selected altogether. All the year three senior secondary school (SSS3) students in all the 30 selected schools formed the sample. The instrument for the study was the 2014 WASSCE multiple-choice test items. Data was analyzed using, Stout's Test of Essential Unidimensionality, BMPAT and Item factor analysis.

Results

Hypothesis one: 2014 WASSCE Mathematics test data is essentially unidimensional.

To test this hypothesis, Stout's Test of Essential Unidimensionality implemented in DIMTEST was used. Furthermore, BMPAT for undimensionality test was also used for cross-validation of the result obtained from the DIMTEST. To conduct the BMPAT, unidimTest one of the subsidiary of Itm, an R package was used. The results are presented as follow:

Unidimensionality test under DIMTEST

To perform the test, the items were divided into two subtests that are as dimensionally distinct as possible, the Partitioning Subtest (PT)and the Assessment Subtest (AT). Items that might form a secondary dimension, the Assessment Subtest, were selected empirically, using the HCA/CCPROX cluster procedure and DETECT statistic in DIMTEST, and this candidate cluster was tested to see if it was dimensionally distinct from the remainder of the test. The null hypothesis is that the responses are unidimensional (the average covariance within groups = 0), so failure to reject the null hypothesis indicates that the assumption of unidimensionality is tenable. Table 2 presents the result

Table 2: Unidimensionality of 2014 WASSCE Mathematics test

TL	TGbar	T	p-value
15.3503	7.9336	7.3799	0.000

Table 2 showed that the AT were dimensionally distinct from each of the remaining items of the test(T = 7.3799 (p-value = 0.00, one-tailed)); therefore, the assumption of unidimensionality was rejected. This showed that there were more than one dimension that accounted for the variation observed in students responses to the Mathematics test items.

Unidimensionality test under BMPAT

To test the assumption of unidmensionality of the Mathematics test using BMPAT, the UnidimTest, a subsidiary of Item an R package was used. To achieve this, the data was simulated using 3PL, the 3PL fitted the data more than the 2PL and 1PL. The result is presented as follows:

Table 3: Bootstrap modified parallel analysis test of unidimensionality

	Value	p-value
second eigenvalue in observed data	3.9285	0.0099
Average of second eigenvalues in montecarlo samples	1.2669	

The result showed that second eigenvalue of the observed data is larger (3.9285) than the second meaneigenvalues of the simulated data (1.2669). Furthermore, BMPAT showed that the observed difference was statistically significant (p = 0.0099). This result showed that the 2014 WASSCE Mathematics test items are not unidimensional. This result matches what was found using the Stout's Test of Essential Unidimensionality. Hence, the hypothesis "2014 WASSCE Mathematics test is essentially unidimensional" was rejected. These results indicated that WASSCE 2014 Mathematics is multidimensional

Research question 1

How many traits underlying WASSCE 2014 Mathematics test account for variation in examinees responses?

To answer this research question, item factor analysis was conducted. For a start, two and three factor model were hypothesized for the data, follow by three and four factor model and so on and in turn the fitness of the model were compared using AIC, BIC, SBIC, and Likelihood ratio test. The factor model with the best fit provided the information for the number of dimensions underlying the data set. These analyses were all conducted using MIRT package of R programming language. The results are presented a follow:

Table 4: Dimensionality of 2014 WASSCE Mathematics multiple choice test

No. of dimension	AIC	AICc	SABIC	BIC	logLik	X2	df	p
two and three-dimension models compared								
2	65067.7	65152.2	65438.68	66070.76	-32334.85	1485.918	48	0.000
3	63677.78	63814.82	64138.24	64922.79	-31591.89			
three and four-dimension models compared								
3	63677.78	63814.82	64138.24	64922.79	-31591.89	1160.921	47	0.000
4	62610.86	62815.65	63158.94	64092.78	-31011.43			
four and five-dimension models compared								
4	62610.86	62815.65	63158.94	64092.78	-31011.43	439.486	46	0.000
5	62263.37	62552.86	62897.21	63977.15	-30791.69			
five and six-dimension models compared								
5	62263.37	62552.86	62897.21	63977.15	-30791.69	138.029	45	0.000
6	62215.34	62608.49	62933.07	64155.95	-30722.67			

Table 4 shows that when two and three dimensions were hypothesized to underlie the 2014 WASSCE Mathematics test, the result showed that the three dimension-model had AIC, AICc, SABIC and BIC values were lesser than the AIC, AICc, SABIC and BIC values of the two-dimension model . In addition, the Likelihood ratio was statistically significant ($\chi^2(48) = 1109$, $p < 0.005$). These results showed that the three-dimension model fitted the data better than the two-dimension model. In search for a better fit for the test data, the three-dimension model was in turn compared with four-dimension model. The result showed that the four-dimension model fitted the data better than the three-dimension model (four-dimension model's AIC, AICc, SABIC and BIC values were respectively lesser than the three-dimension model's AIC, AICc, SABIC and BIC; the Likelihood ratio was statistically significant, ($\chi^2(48) = 1109$, $p < 0.005$)).

In the same vein, the model-data fit forfour-dimension model and five-dimension model were compared and the results showed that the five-dimension model fitted the data better than the four-dimension model (five-dimension model's AIC, AICc, SABIC and BIC values were lesser

than the four-dimension model's AIC, AICc, SABIC and BIC values respectively; Likelihood ratio was statistically significant, χ^2 (48) = 1109, p < 0.005). Furthermore, the fitness of the five-dimension model to the data was compared to that of six-dimension model and the results showed a mixed picture.The six-dimension model fitted the data better than the five-dimension model based on the likelihood ratio test (p < 0.005). However, the information indices, AICc, SABIC and BIC indicated that five-dimension model fit the data better (lower values).

These results showed that five-dimension model clearly fitted the data without any controversy when the four-dimension model. Although, the six-dimension model also appeared to fit the data, there were conflicting issues in the analysis of the fit indices. While some of the indices favoured five-dimension model, others favoured the six-dimension model. Based on this and the application of Occam's razor,we concluded that the five-dimension model fitted the data.

These results showed that there were five dimensions that underlie the 2014 WASSCE Mathematics test

Research question 2: What is the nature of the 2014 WASSCE Mathematics dimensionality?

To answer this research question, Full information factor analysis based on the optimal model (five-dimension model) was conducted. The factor loadings resulting from the factor analysis were used in assessing the nature of the test's dimensions. The results of the factor loadings are presented as follow:

Table 5: 2014 WASSCE Mathematics multiple choice test dimensionality nature

item	F1	F2	F3	F4	F5
Item1	**-0.89735**	-0.0316	-0.2455	-0.09162	0.13632
Item2	**-0.75908**	-0.1803	-0.1457	-0.22471	-0.00498
Item3	**-0.9288**	0.0138	0.0826	-0.11442	0.15678
Item4	**-0.91912**	-0.0318	-0.2195	-0.02429	0.04683
Item5	**-0.92776**	-0.1399	-0.1293	-0.06441	-0.09385
Item6	**-0.86483**	0.1023	-0.0787	-0.07624	0.11886
Item7	**-0.89071**	-0.1699	**0.3187**	0.25172	-0.34389
Item8	**-0.87479**	0.1053	0.0646	-0.02421	0.0912
Item9	**-0.95375**	0.0762	0.0569	0.0334	0.0089
Item10	**-0.88488**	0.1187	0.1243	0.16317	0.06625
Item11	**-0.57263**	0.2962	-0.1933	**0.38887**	0.14195
Item12	**-0.37868**	0.0295	0.1714	**0.82907**	0.05014
Item13	**-0.78305**	-0.0854	-0.1103	**0.33816**	0.14557
Item14	**-0.43437**	0.1644	-0.0766	**0.64399**	**0.35942**
Item15	**-0.36009**	-0.032	0.1893	**0.5743**	**0.49096**
Item16	-0.26399	-0.0457	0.0561	0.06027	**0.75219**
Item17	0.00359	-0.2397	0.314	**0.77207**	0.24723
Item18	**-0.42408**	0.0611	0.0118	0.23212	**0.66337**
Item19	**-0.32092**	-0.0203	-0.085	-0.10499	**0.69405**
Item20	-0.085	**-0.5446**	**0.4323**	**0.36005**	0.02949
Item21	**-0.53365**	0.0408	0.086	0.18739	**0.47995**
Item22	**-0.72828**	0.2068	0.0871	0.11287	0.17433
Item23	**-0.40474**	**0.7379**	0.267	-0.0902	0.04234
Item24	**-0.4058**	0.2861	-0.0168	**-0.5643**	0.10485

Item25	**0.35311**	0.1438	-0.0423	**0.92599**	-0.18369
Item26	-0.29589	0.1854	**-0.4682**	**0.78584**	-0.08329
Item27	-0.12519	0.684	**0.5085**	-0.12359	-0.22726
Item28	0.22983	-0.1102	-0.0716	0.10871	-0.01827
Item29	0.01907	-0.1349	-0.3038	**0.56775**	-0.6675
Item30	**-0.53119**	0.2466	0.2038	**-0.51713**	0.0674
Item31	-0.05705	**-0.4368**	**0.4013**	-0.13516	-0.02539
Item32	-0.20969	-0.1099	0.1487	-0.17193	0.14543
Item33	0.16248	**1.0266**	-0.2226	0.18398	0.13771
Item34	-0.25003	-0.1761	**0.4428**	**0.65929**	-0.25781
Item35	-0.08493	**0.5978**	**0.5591**	0.00496	**-0.42792**
Item36	-0.00609	0.1228	**0.9666**	0.09191	-0.14564
Item37	-0.04778	**0.5787**	**0.716**	-0.04529	-0.05769
Item38	-0.0335	0.1816	**0.6012**	-0.257	**0.3728**
Item39	-0.18144	0.182	**0.8205**	-0.15599	0.17934
Item40	0.0913	0.0224	**0.8284**	0.28734	**-0.32236**
Item41	-0.03778	0.2838	**0.7738**	**-0.42246**	-0.22894
Item42	0.20325	-0.1752	**0.9749**	0.13535	-0.07804
Item43	-0.12104	-0.1965	**0.9826**	0.09076	-0.04606
Item44	0.11995	0.0693	**0.8456**	-0.04608	**0.3308**
Item45	-0.19393	0.2137	**0.5922**	-0.1129	0.22305
Item46	0.08954	0.1215	**0.7705**	-0.08464	**0.4094**
Item47	-0.20402	-0.0795	**0.3682**	0.15633	0.28242
Item48	0.03684	-0.2082	**0.8774**	-0.04398	**0.3639**
Item49	-0.2855	0.1065	0.2852	0.02931	0.27072
Item50	-0.30594	0.0773	0.2728	0.06882	0.2782

Bold faced loadings are loadings with values greater than or equal to 0.32 (substantial loading); items having more than one substantial loading showed multidimensionality.

Table 5 presents the unsorted factor loadings of the Mathematics multiple choice test after rotation. The table shows that the five factors (F1, F2, F3, F4, and F5) out of the six have more than three loadings greater than or equal to 0.32, the condition set for adjudging a factor well defined (Tabachnick & Fidell, 2013). Precisely, the table shows that items (1-6, 8-10, and 22) loaded on F1; only item 33 loaded on F2. On F3 item 27, 36, 42, 43, 45 and 47 were loaded; on F4 two items: 17 and 29 were loaded and on F5 only one item, item 16 was loaded. Furthermore, some items loaded on more than one factors. Items in this category include: 7(loaded on F1 and F3); 11, 12, 13, 24, 25 and 30 (loaded on F1 and F4); 31 and 37 (loaded on F2 and F3); 23 (F1 and F2); 18, 19, and 21 (loaded on F1 and F5); 26, 34, 41, and 46 (loaded on F3 and F4); 38, 40, 44 and 48 (loaded on F3 and F5); 14 and 15 (loaded on F1, F4 and F5); 35 (loaded on F2, F3 and F5) and 20 (loaded on F2, F3 and F4). These results showed that the test was multidimensional and the observed cross-loading showed evidence within-item multidimensionality.

Discussions

The results revealed that the 2014 WASSCE violates the unidimensionality assumption implicit in the classical test theory, a measurement framework adopted by the West African Examination Council for the scoring of the test. This finding negates the finding of the study of Awopetu and

Afolabi (2016). The authors concluded that 2011 SSCE Mathematics developed by the National Examinations Council was unidimensional. The difference observed in the present study and that of Awopetu and Afolabi (2016) could be as a result of the statistical tools employed in the studies. In Awopetu and Afolabi study, exploratory factor analysis implemented in SPSS, a factor analysis developed for continuous data, was used. In the present study, item factor analysis (full information factor analysis), a factor analysis designed for factor analyzing item responses that are dichotomously scored was used. The implication of the findings of the present study is that only one of the five traits being measured by the Mathematics test was assessed. This is because the measurement framework, CTT used for the estimation of examinees' test score can only account for one trait and any other traits underlie the test are considered as nuisance traits and are treated as error. Thus, the true performance of the examinees in the test may be compromised.

Conclusion and Recommendation

This study sought to assess the appropriateness of using CTT for the scoring of examinees' test performance in the 2014 WASSCE Mathematics test. To achieve, the unidimensionality, a necessary condition that must be fulfilled by the Mathematics test data before it can be subjected to CTT scoring, was tested. Based on the results obtained in the course of data analysis, it is our conclusion that the 2014 WASSCE Mathematics test violated the assumption of unidimensionality. In fact, there were five dimensions which underlie the test data. While some of the items measure only one trait, some measure two trait and some three traits. Hence, the study recommended that unidimensionality of Mathematics tests should not be assumed; the assumption should be tested and the number of dimensions should be assessed prior to selecting measurement framework for the estimation of test scores.

References

Ackerman, T.A., Gieri, M.J., Walker, C.M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests.*Educational Measurement: Issues and Practice* 22: 37 -51

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice* 4.22: 87 – 96.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In b.n. Petrov& f. Caski (eds), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–81). budapest: akademiai Kiado.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C: American Educational Research Association.

Ansley, R.A., &Forsyth,T.N. (1985). an examination of the characteristics of unidimensionalIrt parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37–48.

Aremu, O.A. & Sokan, B.O. (2003). A multi-casual evaluation of academic performance of Nigerian learners: issues and implications for national development. Department of Guidance and Counselling, University of Ibadan, Ibadan.

Awopeju, O.A.&Afolabi, E.R.I. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an eM algorithm. *Psychometrika*, 46(4), 443–59.

Bolaji, C. (2005). A study of factors influencing students' attitude towards mathematics in the Junior Secondary Schools; Mathematics teaching in Nigeria. Retrieved on March 25 2010 from http://www.2.ncsu.edu/ncsu/aern/bolajim.html

Budescu, D.V., Cohen, Y., & Ben-simon, A. (1997).a revised modifid parallel analysis for the construction of unidimensional item pools. *Applied Psychological Measurement*, 21(3), 233–52.

Cai, l. (2010). High-dimensional exploratory item factor analysis by a Metropolis–hastings robbins–Monro algorithm. *Psychometrika*, 75, 33–57.

De Ayala, R.J.(2009). *The theory and practice of Item Response Theory*. New York: The Guilford Press

Demars, C. (2010). *Item response theory: Understanding Statistics Measurement*. Oxford: Oxford University Press

Edwards, M.C,Wirth,R.J, Houts, C.R.&Bodine, A.J. (2015). Three (or four) factors, four (or three) models. In Reise, S.P and Revicki, D.A. (Eds) *Handbook of Item Response Theory Modelling: Applications to typical performance assessment.* London: Routledge

Finch, W.H. & Brian, B.F .(2015). Latent Variable Modeling with R. Routledge: London

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modifid parallel analysis for Irt dimensionality assessment. *Applied Measurement in Education*, 21, 119–40.

Fraser, C. & McDonald, R.P. (2012).NOHARM 4.0.A Windows programme for fitting both unidimensional and multidimensional normal ogive models of latent traits theory [Computer Programme].Welland, ON: Niagara College.

Garrido, L.E. Abad, F.J.&Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables.*Psychological Methods*. 18.4: 454-474

Gorsuch, R.L. (1983). *Factor analysis.*Hillsdale, NJ: Lawrence Erlbaum

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

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139–64.

Henson, R.K.& Roberts, J.K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement*. 66: 393 - 416

Horn.(1965). A rationale and test for the number of factors in factor analysis. *Psychometrika.*30: 179-185

Joreskog, K.G. &Sorborm, D.(2006). LISREL 8.80 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Kaiser, H.F.(1960). The application of electronic computers to factor analysis.*Educational and Psychological Measurement,* 20: 141 - 151

Kline, T. J. (2005). Classical test theory assumptions, equations, limitations, and item analyses.*Psychological Testing.* T.J. Kline. Calgary, Canada: SAGE Publications.

Ledesma, D& Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis.*Practical Assessment, Research and Evaluation*.12. http://pareonline.net/pdf/v12n2.pdf.Retrieved October, 20th 2016.

Lorenzo-Seva, U.&Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behaviour Research Methods.*38: 88 -91

Schwarz, G.E. (1978). Estimating the dimension of a model.*Annals of Statistics* 6, 461–4.

McDonald, R.P. (1999). *Test theory: a unified treatment.* Hillsdale NJ: Erlbaum

McDonald, R.P. (1985). *Factor analysis and related methods.* Hillsdale NJ: Erlbaum

Metibemu, M.A. (2016) Comparison of Classical Test and Item Response Theories in the Development and Scoring of Senior Secondary School Physics Tests in Ondo State, Nigeria. Unpublished PhD Thesis, Institute of Education, University of Ibadan, Ibadan.

Muthen, L.K.&Muthen, R. (2014). Mplus Version 7.4 [Computer Software] Los Angeles , CA ;Muthen&Muthen

National Mathematics Centre, Abuja .(2009).Mathematics improvement Programme.www.nmcabuja.org/mathematics_improvement_programmes.html. Retrieved 26th July, 2010.

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

Peres-Neto, P. Jackson, D.& Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis.*49: 974- 997

Reckase.M.d. (1985).The diffiulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401–12.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986).The interpretation of unidimensionalIrt parameters when estimated from multidimensional data. Paper presented at the annual Meeting of the Psychometric society.

Reckase, M.D. (2009). *Multidimensional item response theory.* London: Springer

Enders, C.K., &Tofihi, D. (2008).the impact of misspecifying class-specifi residual variances in growth mixture models. *Structural EquationModeling: A Multidisciplinary Journal*, 15, 75–95.

Stout, W. (1987).A nonparametric approach for assessing latent trait dimensionality.*Psychometrika*52: 589 - 617

Stout, W. (2005). DIMTEST (Version 2.0) [Computer Software]. Champaign. IL: The William Stout Institute for Measurement.

Tabachnick, B. G., &Fidell, L. S. (2007).*Using Multivariate Statistics*. Boston: Pearson

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items.*Applied Psychological Measurement. 27.3: 159 - 202*

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations.*Psychometrika*. 41: 321 - 327

Yu,C.H., Popp S.O., DiGangi, S.&Jannasch-Pennel, A. (2007). Assessing unidimensionality: A comparison of Rasch modeling, parallel analysis and TETRAD.*Practical Assessment, Research and Evaluation.*12.14. http://pareonline.net/getvn.asp?v=12&n=14

Wang, W. C. (1995). Implementation and application of the multidimensional random coefficients multinomial logit.Unpublished doctoral dissertation.University of California, Berkeley.