**Transactions of the Nigerian Association of Mathematical Physics**
**Vol. 6 (Jan., 2018), pp. 344 - 354**


***In honour of Prof. Ekhaguere at 70***

# Unidimensionality and local independence assumptions of item response theory of the WAEC and NECO standardised Mathematics tests

**P. O. Olonade [a] and J. G. Adewale[b]**

[a,b]*International Centre for Educational Evaluation, Institute of Education, University of Ibadan, Ibadan, Nigeria*

**Abstract.** This study investigated the assumptions of Item Response Theory (IRT) in the standardised mathematics tests of Senior School Certificate Examination (SSCE) of the WAEC and the NECO in Lagos State using factor analysis and polychoric correlation methods. The 50-item 2014 SSCE Standardised Multiple Choice Mathematics Tests from each of WAEC and NECO were administered on two thousand, one hundred and forty nine (2,149) students randomly selected from 48 schools in Lagos State, Nigeria. The responses of the students were analysed using factor analysis, polychoric correlation. Results showed that the 50-item each of WAEC and NECO 2014 SSCE measure students' mathematics ability establishing unidimensionality of the tests; none of the items provided clue for answering another item establishing local independence of the test items. Therefore, unidimensionality and local independence assumptions should be obeyed in assessment that uses IRT.

**Keywords:** Item response theory, unidimensionality, local independence, standardised Mathematics tests.

## 1. Introduction

Students are always tested in schools to produce scores which are often used for assessment and taking decisions on promotion, certification, placement, admission and employment. In Nigeria, public examining bodies administer standardised tests on students to assess their abilities, achievement, skills and performance. They obtain their standardised tests by developing test items, pilot testing, calibrating and coding their test items, banking the items and embarking on other necessary processes. The administered tests are marked and the overall scores are used to assess any character trait of the students without paying attention to the response pattern of the students to each test item in relation to the students' character traits. Assessment is expected to improve the fairness of the items of any measuring instrument if appropriate decisions are to be made on the outcome of the tests. According to Roever (2005), a fair test is one which affords all the testees equal opportunity to demonstrate the skills and knowledge which they have acquired and which are relevant to the purpose of the test. A fair test should contain items that are easy for students to respond to. When two students answer different items that have different difficulty level, the students are not to have the same score value because ability level used to answer different items is not the same. In fairness to student that answer more difficult question, the score should be higher. Teachers, examiners and researchers should focus on improving the fairness of the test items such that testees will be able to give correct response to the test items. If the students' ability in responding to any test item is not considered, their achievement in such test will not produce a desired result. A theory that enables attention to be paid to improving individual test item is IRT. IRT has been developed to

provide a framework for evaluating how well assessments work, and how well individual items on assessments work.

Evidence from literature shows that assessment of students based on their pattern of response to test items concentrated on developing tests, validation and calibration (Enu, 2015), banking of items (Akindele, 2004), comparability of test items parameter estimates between Classical test theory (CTT) and Item response theory (IRT) models (Adedoyin and Adedoyin, 2013) and equating the difficulties for two versions of exams (that allows comparison between testees' scores). Some of these studies investigated the assumption of IRT in many subject areas including mathematics. This therefore necessitated a comparison of two forms of SSCE examination that SS3 students enrol for every year conducted by two prominent examining bodies in Nigeria, the West African Examinations Council (WAEC) and National Examinations Council (NECO). This is to ensure that they can be used for assessment under Item response theory (IRT) framework.

Some of these assessment practices have been sustained by two theoretical frameworks which are classical test theory (CTT) and item response theory (IRT). The classical test theory has been the foundation for measurement theory for over 80years(Allen and Yen, 2002; Hambleton and Jones,1993), yet it is faced with the problems of non-correlation of true and error scores, group dependence item statistics, assumption of equal errors of measurement among all testees (Enu, 2015). This gave rise to the development of Item Response Theory (IRT) generally claimed as improvement over CTT.IRT is a set of models which, by relating the likelihood of a particular reaction by an individual with a given trait level to the characteristics of the item designed to elicit the level to which the individual possesses that trait (Nenty, 2004; Rupp, 2009).IRT is used to estimate the parameters and testees' unobserved traits such that there is an encounter or interraction between individual testee and an item during the testing process; hence IRT is regarded as latent trait theory which focuses on the test items, unlike CTT that focuses on test scores.

Any assessment meets certain basic IRT assumptions which include: 1.Unidimensionality of trait denoted by $\Theta$;2. Local independence of the items; 3.The response of a person to an item can be modeled by a mathematical item response function (IRF).A unidimensional IRT model should ensures that a single underlying construct or trait (e.g. ability level) of each testee is measured by the items. Items should also be locally independent such that the probability of solving any item by a testee is independent of the outcome of any other items, controlling for latent ability levels and item parameters. Lastly, the response of a testee to an item can be modeled by a mathematical item response function (IRF). In this case, each item on a test has its own item characteristic curve that describes the probability of getting each particular item right or wrong given the ability level of each test taker. Example of mathematical IRF is the three parameter logistic (3PL) model defined as:

$$p_i(\theta) = c_i + \frac{1-c_i}{1+e^{-a_i(\theta-b_i)}} \qquad (1)$$

which indicates the probability that an examinee with ability $\theta$ will respond correctly to a dichotomous item *i*, usually a multiple-choice question, is given by item discrimination parameter a, difficulty parameter b, and guess factor c, where $\theta$ is modeled as a sample from a normal distribution for the purpose of estimating the item parameters. Other unidimensional models used for multiple-choice format of objective test forms are two-parameter logistic model (2PL) and one-parameter logistic model (1PL).

Assessment using IRT by public examining bodies like WAEC and NECO that conducts SSCE is desirable. WAEC, since inception has been confronted with the challenges of producing small number of candidates obtaining credit pass in the senior secondary school certificate examination result, especially in mathematics. After many years of its existence, the Nigerian

government came up with NECO to serve as alternative examining body. The same story repeats itself every year that students are not able to obtain high score in mathematics probably because the questions (items) are difficult to solve on one hand and the students are not able to solve them on the other hand. It was observed during field work that some schools in Lagos state don't enroll their students for NECO as the yearly result scores produced indicate that NECO is still assumed to be more difficult than WAEC.

Therefore, mathematics examination forms of SSCE conducted by these two organizations need to be compared because of the importance of the subject in the placement of students for the desired course in the higher institution. Some comparison made in the pastby researchers in many developing countries still usetraditional methods for the fundamental aspects of assessment process like scoring and calibration methods. According to Adegoke (2015), most public examining bodies such as WAEC, NECO and NABTEB use number- correct in the estimation of the ability scores of their candidates. This is due to the fact that the use of IRT appears to be more technical than Classical Test Theory (CTT) framework.Nigerian based examining bodies are expected to use IRT for assessment.

Using IRT method, parameters are estimated either by concurrent item calibration (Wingersky & Lord, 1984) or separate calibration. Concurrent calibration is accomplished using certain computer programs, such as LOGIST (Wingersky, 1983), MULTILOG (Thissen, 1991) or BILOG-MG (Zimowski et al., 1996)which can undergo a single calibration run to estimate item parameters simultaneously. This also enables the item parameter estimates for the two test forms to be equated automatically beside other assessments. This is because all estimated parameters are on the same scale and transformation of item parameters is automatically performed. Separate calibration (estimation) of item and ability parameters is accomplished by BILOG-MG software package when item parameters are estimated separately for each test form, the units of the item parameters are not on the same scale because two groups of examinees differ in ability levels; so their mean ability levels and the standard deviations are not equal. The scale for measuring ability is determined up to an arbitrary linear transformation. Typically, this indeterminacy is solved in such a way that the mean and standard deviation of ability parameters are arbitrarily fixed to 0 and 1 ("0, 1"scale) for the group of examinees at hand. Scaling can therefore be described as a process of transforming student raw scores (i.e., number correct) onto a different score scale so as to facilitate interpretation and understanding of test scores. Pang, Madera, Radwan and Zhang (2010) opine that the students' ability scores for two tests are estimated separately using the corresponding scaled item parameters, and the means of the two tests are then compared to determine the direction and magnitude of the change.

The importance of establishing the assumptions of IRT in assessment of any students' trait in SSCE mathematics cannot be overemphasized. If assessment is to be successfully carried out, the items of the test must measure exactly the traits (e.g. ability) for which they are developed and the trait should be able to explain examinees' performance in the test. Similarly, items need to be so examined to ensure that one item does not suggest an answer to another item in the test. Though, SSCE mathematics of WAEC and NECO are standardised tests, checking these assumptions is very necessary because of underlining assessment to be carried out. The study that prompted establishing these assumptions sought to assess the students' abilities using IRT methods to equate the scores obtained in WAEC and NECO mathematics objective tests. Since measuring ability becomes our concern, we need to assume that the two tests measure only one trait (students' ability) on a standard scale (with a mean of 0.0 and standard deviation of 1.0) and that the items in the tests are not related except that they measure exactly the same trait. Hence, the study assumes unidimensionality of the tests and local independence of the items.

### *1.1 Research questions*

The study seeks to provide answers to the following questions:

1. To what extent do the 50-item mathematics of WAEC and 50-item mathematics of NECO obey the assumption of unidimensionality under Item Response Theory (IRT) framework?
2. To what extent do the 50-item mathematics of WAEC and 50-item mathematics of NECO obey the assumption of item local independence under Item Response Theory (IRT) framework?

## 2. Methodology

The study adopted the single-group with counterbalance equating design. The study was carried out among the senior secondary three (SS3) students that registered for mathematics in senior school certificate examination (SSCE) of WAEC and NECO in Lagos State, Nigeria. The state has twenty local government areas and was stratified along the six educational districts. Simple random sampling was used to select a local government area from each educational district and eight schools from each local government area. The total number of schools selected is 48. An intact class was randomly selected from each of the schools. The sample size was 2,149 of SS3 mathematics candidates of WAEC and NECO.

The instruments used for data collection were2014 SSCE Mathematics WAEC Multiple Choice Test (SWMT)-types $W_1$ & $W_2$ and 2014 SSCE Mathematics NECO Multiple Choice Test (SNMT)-types$N_1$& $N_2$.The standardised WAEC test consists of 50 items each having four response options. The test was used as type $W_1$. $W_1$ items were rearranged to form type $W_2$. Similarly, standardised NECO test consists of 50 items drawn from original paper2 (containing 60 items) of SSCE (type$N_1$) with each item having five response options after **removing 10 items** and rearranging them to have the same position in terms of content with $W_1$ items of WAEC test. **Type $N_2$**was formed by rearranging type $N_1$ items. The instruments were administered on the total of 2,149 SS3Mathematics students in groups 1&2. Type $W_2$and $N_2$ items were reshuffled back to original type $W_1$ and$N_1$ items position respectively for the purpose of analysis. In order to achieve successful item calibration and subsequent score equating, a necessary condition of equal test length of 50items of WAEC and NECO mathematics must be met, hence 10items were removed from 60 items of original NECO mathematics.

Data collection was done in November and December, 2015 at the eve of revision and terminal examination. The official letters of introduction were collected from the Institute of Education andLocal Educational District, taken to the school administrators for permission to administer the tests to their students. Tests were administered to the two groups for 3hrs (1 ½ hrs for each paper) with the help of trained research assistants and cooperating teachers in some schools according to the chosen equating design where $W_1$ and $N_1$ were taken by group1 and $N_2$ and $W_2$ by group 2 for the 1st and 2nd duration respectively. The responses of the examinees for WAEC and NECO forms of SSCE were collected.

Factor analysis was used to establish the unidimensionality of the tests. Polychoric correlation matrices were generated from the data to establish local independence of test items. BILOG-MG program separately run the data with 2PL model that fit the data to calibrate two item parameters and estimate examinees ability into two data files (NECO and WAEC) using marginal maximum likelihood estimation method. 2-parameter logistic (2PL) model used for estimation is expressed as:

$$p_i(\theta) = \frac{1}{1+e^{-a_i(\theta-b_i)}} \qquad (2)$$

## 3. Results

**Research Question 1**: To what extent have the 50-item mathematics of WAEC and 50-item mathematics of NECO obeyed the unidimensionality assumption of IRT framework?
Table 1A, B present statistics of the unidimensionality of the 50 items contained in each of the WAEC and NECO questions through Total Variance Explained.

From Table 4.1A, the highest eigenvalue for WAEC mathematics is 4.59 and is for component one. This shows that the largest component explains 9.18 % of the variance. Although this value is relatively small, it explains that the majority of the items contained in the test hang together on one distinct factor. Also, eigenvalues equal to or greater than 1.00 were extracted. Out of 50 items used as variables 15 were extracted with cumulative variance for all sum of square loading estimated as 44.987 percent. This indicates the extent of the unidimensional trait is about 45% of what makes the WAEC mathematics items valid.

Table 1A :Total variance explained by factor analysis of WAEC mathematics test

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 4.588 | 9.176 | 9.176 | 4.588 | 9.176 | 9.176 |
| 2 | 2.730 | 5.459 | 14.636 | 2.730 | 5.459 | 14.636 |
| 3 | 1.469 | 2.938 | 17.574 | 1.469 | 2.938 | 17.574 |
| 4 | 1.436 | 2.872 | 20.446 | 1.436 | 2.872 | 20.446 |
| 5 | 1.287 | 2.574 | 23.020 | 1.287 | 2.574 | 23.020 |
| 6 | 1.181 | 2.362 | 25.382 | 1.181 | 2.362 | 25.382 |
| 7 | 1.177 | 2.354 | 27.736 | 1.177 | 2.354 | 27.736 |
| 8 | 1.147 | 2.293 | 30.029 | 1.147 | 2.293 | 30.029 |
| 9 | 1.127 | 2.255 | 32.283 | 1.127 | 2.255 | 32.283 |
| 10 | 1.107 | 2.215 | 34.498 | 1.107 | 2.215 | 34.498 |
| 11 | 1.096 | 2.191 | 36.690 | 1.096 | 2.191 | 36.690 |
| 12 | 1.058 | 2.116 | 38.806 | 1.058 | 2.116 | 38.806 |
| 13 | 1.054 | 2.108 | 40.915 | 1.054 | 2.108 | 40.915 |
| 14 | 1.026 | 2.052 | 42.967 | 1.026 | 2.052 | 42.967 |
| 15 | 1.010 | 2.020 | 44.987 | 1.010 | 2.020 | 44.987 |
| 16 | .996 | 1.991 | 46.978 | | | |
| + | + | + | + | | | |
| + | + | + | + | | | |
| 46 | .619 | 1.238 | 95.581 | | | |
| 47 | .607 | 1.215 | 96.796 | | | |
| 48 | .573 | 1.146 | 97.942 | | | |
| 49 | .548 | 1.097 | 99.039 | | | |
| 50 | .481 | .961 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

Similarly, From Table 1B, the highest eigenvalue for NECO mathematics is 3.76 and is for component one. This shows that the largest component explains 7.52 % of the variance. Although this value is relatively small, it explains that the majority of the items contained in the test hang

together on one distinct factor. Also, eigenvalues equal to or greater than 1.00 were extracted. Out of 50 items used as variables 18 were extracted with cumulative variance for all sum of square loading estimated as 49.06 percent. This indicates that the extent of the unidimensional trait is about 49% of what makes the NECO mathematics items valid.

Table 1B :Total variance explained by factor analysis of NECO mathematics test

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3.760 | 7.520 | 7.520 | 3.760 | 7.520 | 7.520 |
| 2 | 2.670 | 5.340 | 12.860 | 2.670 | 5.340 | 12.860 |
| 3 | 1.344 | 2.688 | 15.548 | 1.344 | 2.688 | 15.548 |
| 4 | 1.295 | 2.589 | 18.138 | 1.295 | 2.589 | 18.138 |
| 5 | 1.238 | 2.475 | 20.613 | 1.238 | 2.475 | 20.613 |
| 6 | 1.215 | 2.431 | 23.044 | 1.215 | 2.431 | 23.044 |
| 7 | 1.186 | 2.373 | 25.416 | 1.186 | 2.373 | 25.416 |
| 8 | 1.157 | 2.314 | 27.730 | 1.157 | 2.314 | 27.730 |
| 9 | 1.153 | 2.306 | 30.036 | 1.153 | 2.306 | 30.036 |
| 10 | 1.112 | 2.225 | 32.261 | 1.112 | 2.225 | 32.261 |
| 11 | 1.089 | 2.179 | 34.439 | 1.089 | 2.179 | 34.439 |
| 12 | 1.083 | 2.167 | 36.606 | 1.083 | 2.167 | 36.606 |
| 13 | 1.076 | 2.152 | 38.759 | 1.076 | 2.152 | 38.759 |
| 14 | 1.056 | 2.111 | 40.870 | 1.056 | 2.111 | 40.870 |
| 15 | 1.045 | 2.090 | 42.960 | 1.045 | 2.090 | 42.960 |
| 16 | 1.030 | 2.060 | 45.020 | 1.030 | 2.060 | 45.020 |
| 17 | 1.019 | 2.039 | 47.059 | 1.019 | 2.039 | 47.059 |
| 18 | 1.001 | 2.002 | 49.060 | 1.001 | 2.002 | 49.060 |
| 19 | .983 | 1.966 | 51.027 | | | |
| + | + | + | + | | | |
| + | + | + | + | | | |
| 48 | .597 | 1.193 | 97.944 | | | |
| 49 | .579 | 1.158 | 99.102 | | | |
| 50 | .449 | .898 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

The results of this factor analysis revealed that the 50-item WAEC and 50-item NECO tests are unidimensional. The extent of unidimensionality explained for WAEC and NECO mathematics are not the same as found in the study of Aliyu (2015) which was almost 100%. Since the first eigenvalue of WAEC was 4.588 which is greater than the next 14 eigenvalues and also explained 9.176% of the variance in the dataset; and the first eigenvalue of NECO was 3.76 which is greater than the next 17 eigenvalues and also explained 7.52% of the variance in the dataset, then WAEC and NECO mathematics are unidimensional according to Svend and Christiensen (2002) who pointed out that "what is required for the unidimensionality assumption to be met adequately is the presence of the dominant factor that influences test performance". Also, according to McBride and Weiss (1974), dichotomous test items are unidimensional when the first factor loading for all items is significantly

greater than 1. Orlando, Sherbonve and Thissen (2000) considered dichotomized test items to be unidimensional when the first eigenvalue is substantially greater than the next.

**Research Question 2:** To what extent do the 50-item mathematics WAEC and 50-item mathematics NECO obey the item local independence assumption of IRT framework?
Table 4.1C and D present statistics of polychoric correlations (edited) among the 50 items each of WAEC and NECO mathematics tests.

Table 4.1C: Inter-correlation matrix among the 50 Items of WAEC mathematics

| Items | wi1 | wi2 | wi3 | wi4 | wi5 | + | wi46 | wi47 | wi48 | wi49 | wi50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **wi1** | 1.000 | | | | | | | | | | |
| **wi2** | 0.704 | 1.000 | | | | | | | | | |
| **wi3** | -0.953 | -0.935 | 1.000 | | | | | | | | |
| **wi4** | -0.922 | -0.899 | 0.425 | 1.000 | | | | | | | |
| **wi5** | 0.388 | 0.390 | -0.270 | -0.302 | 1.000 | | | | | | |
| + | + | + | + | + | + | + | | | | | |
| **wi46** | 0.207 | 0.143 | -0.009 | 0.005 | 0.131 | + | 1.000 | | | | |
| **wi47** | 0.051 | 0.061 | 0.117 | 0.113 | 0.057 | + | 0.204 | 1.000 | | | |
| **wi48** | 0.059 | 0.073 | 0.048 | 0.136 | 0.043 | + | 0.161 | 0.107 | 1.000 | | |
| **wi49** | 0.463 | 0.241 | -0.904 | -0.862 | 0.234 | + | 0.058 | 0.121 | 0.262 | 1.000 | |
| **wi50** | -0.940 | -0.920 | 0.460 | 0.405 | -0.287 | + | 0.004 | 0.089 | -0.001 | -0.886 | 1.000 |

Table 4.1D presents the summary of frequencies of the observed polychoric correlation coefficients among the 50 items of WAEC mathematics. It shows 49 correlations of item wi1 with items wi2, item wi3, item wi4, item wi5, and so on up to item wi50. Among the 49 correlation coefficients, 38 have values that are less than or equal to 0.099; 4 have values between 0.100 and 0.199; wi2 have values between 0.200 and 0.299; 2 have values between 0.300 and 0.399; wi2 have values between 0.400 and 0.499 and wi1 has values that are greater than 0.500. Similarly, item wi2 correlates with items wi3, wi4, wi5 …wi50 (giving 48 correlation coefficients). In all there were 1225 correlation coefficients.

Out of 1225 inter-correlations among the 50 items, 788 (64.33%) have polychoric correlation coefficients equal or less than 0.099; 277 (22.61%) have polychoric correlation coefficient between 0.100 and 0.199; etc. This result showed that the percentage of correlation coefficients of WAEC mathematics items that are close to zero is 86.94%. Since the closer the correlation coefficients are to zero the more the items are locally independent, then it can be inferred that WAEC mathematics items were locally independent to the extent of about 87%.

Table 4.1F presents the summary of frequencies of the observed polychoric correlation coefficients among the 50 items of NECO mathematics. For example, for item ni1, there were 49 correlation coefficients in that it correlates with items ni2, ni3, ni4, ni5 . . . ni50. Among the 49 correlation coefficients, 31 have values that are less than or equal to 0.099; 11 have values between 0.100 and 0.199; 7 have values between 0.200 and 0.299; no values for correlation between 0.300 and 0.399; 0.400 and 0.499 and greater than 0.500. In all there were 1225 correlation coefficients.

Out of 1225 inter-correlations among the 100 items, 841 (68.65%) have polychoric correlation coefficients equal or less than 0.099; 264 (21.55%) have polychoric correlation coefficient between 0.100 and 0.199; etc. This result showed that the percentage of correlation coefficients of NECO mathematics items that are close to zero is 90.20%. Since the closer the correlation coefficients are to

zero the more the items are locally independent, then it can be inferred that NECO mathematics items were locally independent to the extent of 90%.

Table 4.1D: Summary of Polychoric correlation coefficients among the 50-Item WAEC mathematics

| Item | ≤ 0.099 | 0.100 to 0.199 | 0.200 to 0.299 | 0.300 to 0.399 | 0.400 to 0.499 | ≥ 0.500 | No of correlations |
|---|---|---|---|---|---|---|---|
| **wi1- wi50** | 38 | 4 | 2 | 2 | 2 | 1 | **49** |
| **Wi2- wi50** | 37 | 6 | 2 | 2 | - | 1 | **48** |
| **Wi3- wi50** | 23 | 7 | 5 | 6 | 4 | 2 | **47** |
| **Wi4- wi50** | 22 | 11 | 8 | 2 | 3 | - | **46** |
| **Wi5- wi50** | 37 | 5 | 2 | 1 | - | - | **45** |
| + | + | + | + | + | + | + | + |
| **Wi47- wi50** | 1 | 2 | - | | - | - | ***3*** |
| **Wi48- wi50** | 1 | - | 1 | | - | - | ***2*** |
| **Wi49- wi50** | 1 | - | - | | - | - | **1** |
| **Total** | **788** | **277** | **81** | **49** | **21** | **9** | ***1225*** |
| **Percentage** | ***64.33*** | ***22.61*** | **6.61** | **4.00** | **1.71** | ***0.73*** | **100%** |

Table 4.1E: Inter-correlation Matrix among the 50 Items of NECO mathematics

| Items | ni1 | ni2 | ni3 | ni4 | nii5 | + | ni46 | ni47 | ni48 | ni49 | ni50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ni1** | 1.000 | | | | | | | | | | |
| **ni2** | 0.117 | 1.000 | | | | | | | | | |
| **ni3** | -0.131 | 0.029 | 1.000 | | | | | | | | |
| **ni4** | 0.123 | 0.101 | -0.079 | 1.000 | | | | | | | |
| **nii5** | 0.224 | 0.108 | -0.124 | 0.243 | 1.000 | | | | | | |
| + | + | + | + | + | + | + | | | | | |
| **ni46** | -0.034 | -0.048 | -0.005 | 0.092 | -0.026 | + | 1.000 | | | | |
| **ni47** | -0.240 | -0.031 | 0.094 | 0.125 | -0.317 | + | 0.099 | 1.000 | | | |
| **ni48** | -0.186 | -0.109 | 0.060 | -0.032 | -0.418 | + | 0.153 | 0.237 | 1.000 | | |
| **ni49** | 0.019 | 0.011 | -0.070 | 0.116 | -0.018 | + | 0.085 | 0.194 | 0.211 | 1.000 | |
| **ni50** | 0.003 | -0.027 | 0.031 | 0.050 | -0.037 | + | 0.031 | 0.086 | 0.024 | 0.192 | 1.000 |

In summary, about 87% of WAEC mathematics items and 90 % of NECO mathematics items possessed observed polychoric correlation coefficients less than 0.200 which is the minimum yardstick for determining level of local independence. This results of analysis revealed that both WAEC and NECO mathematics items were locally independent as most of the polychoric correlation coefficients of their items were close to zero. The condition which informed this finding is in agreement with the condition used for the assessment of local independent assumption of physics pre-test items in Ojerinde (2013) study. In the study, Ojerinde grouped the polychoric correlation coefficient of physics test items into five groups, He concluded that the items were locally independent on the ground that 64.35% of the pairs of items correlation coefficient fell into group one and two. Also, the findings comply with the condition set by Lord (1978) which stipulates that the polychoric correlation coefficients of items obtained from the correlation of the items among one another should not be significantly greater than zero.

Table 4.1F: Summary of Polychoric correlation coefficients among the 50- Item NECO mathematics

| Item | ≤ 0.099 | 0.100 to 0.199 | 0.200 to 0.299 | 0.300 to 0.399 | 0.400 to 0.499 | ≥ 0.500 | No of correlations |
|---|---|---|---|---|---|---|---|
| **ni1- ni50** | 31 | 11 | 7 | - | - | - | **49** |
| **ni2- ni50** | 34 | 11 | 3 | - | - | - | **48** |
| **ni3- ni50** | 45 | 2 | - | - | - | - | **47** |
| **ni4- ni50** | 30 | 15 | 1 | - | - | - | **46** |
| **ni5- ni50** | 31 | 10 | 3 | 1 | - | - | **45** |
| + | + | + | + | + | + | + | + |
| **ni47- ni50** | 1 | 1 | 1 | - | - | - | **3** |
| **ni48- ni50** | 1 | - | 1 | - | - | - | **2** |
| **ni49- ni50** | - | 1 | - | - | - | - | **1** |
| **Total** | **841** | **264** | **78** | **34** | **6** | **2** | **1225** |
| **Percentage** | **68.65** | **21.55** | **6.37** | **2.78** | **0.49** | **0.16** | **100** |

## 4. Conclusion

This study focused on investigating the assumptions of using IRT in any assessment of students using standardized WAEC and NECO multiple choice mathematics tests of 2014SSCE. Results indicated that the tests were highly unidimensional and their items were locally independent. WAEC and NECO mathematics showed that they are unidimensional since the first eigenvalue of WAEC was 4.588 which is greater than the next 14 eigenvalues and also explained 9.176% of the variance in the dataset; and the first eigenvalue of NECO was 3.76 which is greater than the next 17 eigenvalues and also explained 7.52% of the variance in the dataset which pointed to the fact that all items in the test loaded on the first factor that is dominant and influences test performance. The first factor loading of WAEC is greater than that of NECO. The extent to which WAEC mathematics test items measured examinees' unidimensionality of trait is about 45% while that of NECO mathematics is about 49%. Though, WAEC showed higher factor loading than that of NECO but test items of WAEC showed lower unidimensionality than that of NECO. Similarly, the extent to which WAEC mathematics test items were locally independent is about 87% while that of NECO mathematics is about 90%.Since the closer the correlation coefficients are to zero the more the items are locally independent, then it can be inferred that more of WAEC mathematics items were locally independent than NECO mathematics items with a percentage difference of about 3%. We can then conclude that WAEC mathematics items are more stable than that of NECO in measuring students' ability. The establishment of these assumptions then necessitated that the two-parameter logistic (2PL) model of IRT that satisfactorily fit the test data was used to estimate parameters before assessment of equating that the study was originally designed for was conducted.

### *4.1 Recommendations*

Based on the findings of the study the following recommendations were made:

1. Public examining bodies should assess the students using standardised tests under IRT framework.
2. Assumptions of IRT should be established before assessment is completed so as to achieve improvement of the item parameters that will enhance better assessment.
3. Psychometricians should develop and test models and software packages that will take care of all areas of assessment procedure.

## References


[1] Adedoyin, O.O.& Adedoyin J.A. 2013. Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters Herald Journal of Education and General Studies Vol. 2 (3), pp 107 - 114.

[2] Adegoke, B. A. 2015. Effects of Scoring Method on Senior Secondary School Students' Ability Scores in Physics Achievement Test. A Paper Presented at the WAEC Monthly Seminar held at WAEC International Office, Lagos.

[3] Akindele,B. P. 2004. The Development of item bank for the Selection Tests into Nigerian Universities: An Exploratory Study. Unpublished PhD Thesis. University of Ibadan

[4] Aliyu, R. T. 2015. Emerging of item Response Theory Models for construction and validation of mathematics achievement test. A paper presented at the 17th Annual National Conference of Association of Educational Researchers and Evaluators of Nigeria. University of Ibadan

[5] Allen, M.J., & Yen, W. M. 2002. *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.

[6] Enu, V.O., 2015. The use of IRT in the Validation and Calibration of Mathematics and Geography Items for Joint Command Schools Promotion Examination in Nigeria. A thesis in the International Centre for Educational Evaluation (ICEE), Institute Of Education in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Ibadan

[7] Hambleton, R. K. & Jones, R.W. 1993. Comparison oo classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice,12(3)*, 3847

[8] McBride, J.R., &Weiss,D.J.1974. A word knowledge item pool for adaptive ability measurement.(Research Rep74-2). Minneapolis: University of Minesota,Department of Psychology, Psychometrics Methods Program, June 1974.(AD781894)

[9] Nenty, J.K. 2004. From Classical Test Theory(CTT) to Item Response Theory (IRT): An Introduction to a Desirable Transition. In O.A. Afemikhe and J.G. Adewale (Eds), Issues in Educational Measurement and Evaluation in Nigeria (*in honour of"* Wole Falayajo) (Chapter 33,,pp371-384). Ibadan, Nigeria: Institute of Education, University of Ibadan.

[10] Ojerinde 'Dibu. 2013. *Classical Test Theory(CTT) versus Item Response Theory (IRT): An evaluation of the comparability ofitem analysis results.* A guest lecture presented at the institute of education. University of Ibadan on 23rd May

[11] Orlando, M., Sherbourne, C. D., Thissen, D. 2000. Summed-score linking using item response theory: Application to depression measurement. Psychol Assessment. 2000, 12: 354-359.

[12] Pang, X., Madera, E., Radwan, N., & Zhang. 2010. A Comparison of Four Test Equating Methods. Report Prepared for the Education Quality and Accountability Office (EQAO)

[13] Roever, C. 2005. "That,s Not Fair!" Fairness, Bias, and Differential Item Functioning in

[14] Language Testing. http://www2.Hawaii.Edu/Brownbag.pdf.retrieved 18-11-2006, University of Hawaii System Website:

[15] Rupp, A.A.2009. Item Response Modelling with BILOG-MG and MULTILOG for

[16] Windows, *International Journal of Testing*. 3.4:365-385

[17] Svend K, Christensen KB (2002). Analysis of local Independence multidimensional in graphical loglinear Rasch Models. Education and Psychological Measurement. 45 (6): 856-865.

[18] Thissen, D. 1991. *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International, Inc.

[19] Wingersky, M. S. 1983. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Application of item response theory* (pp. 45-56). Vancouver, British Columbia: Educational Research Institute of British Columbia.

[20] Wingersky, M. & Lord, F. 1984. An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8(3), 347–364.

[21] Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. 1996. BILOG-MG: Multiple group IRT analysis and test maintenance for binary items. Computer program.