INVARIANCE PERSON ESTIMATE OF BASIC EDUCATION CERTIFICATE EXAMINATION: CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY SCORING PERSPECTIVE

Musa Adekunle Ayanwale and Joshua Oluwatoyin Adeleke
University of Ibadan, Nigeria
and
Titilayo Iyabode Mamadelo
National Examinations Council, Ogun State, Nigeria

Abstract: *A scoring framework that does not reflect true performance of an examinee would ultimately result in an abnormal score. This study assessed invariance person estimates of 2017 Nigerian National Examinations Council Basic Education Certificate Examination Mathematics Multiple Choice using classical test theory (CTT) and item response theory (IRT) scoring frameworks. The study adopted survey design method. Simple random sampling technique was adopted to select 978 subjects (425 males, 553 females, M=12 years) for the study. One research instrument was used. Data were analyzed using descriptive and paired sample t-tests. There was significance difference (t = 2.635, df = 977, p = 0.01) in the overall mean score of CTT (mean=50.70, SD=10.30) and IRT (mean=47.78, SD=8.49). Also, IRT method of scoring produced different test scores for examinees who had the same raw scores under the CTT method. National Examinations Council should be encouraged to shift their paradigm of scoring from CTT to IRT method.*

Keywords: CTT scoring, IRT scoring, basic education examination, mathematics

Introduction

The *Basic Education Certificate Examination* (BECE) is one of the examinations conducted by Nigerian National Examinations Council (NECO) for examinees who will transit from the three years of junior secondary to the senior secondary category. This examination is what is obtainable globally, and Nigeria is no exception from implementing this aspect of educational system. Candidate are deemed to have passed the BECE if they have credit in six subjects including English and mathematics. Mathematics aspect of BECE conducted by National Examinations Council takes two forms. These forms include (Paper I) – a multiple-choice of 60 items containing items on different themes such as number and numeration, algebraic process, geometry and mensuration, and everyday statistics and (Paper II) – the constructed response test items. In this study, emphasis was on the multiple-choice items because of its ability to cover representative samples of the content of interest without necessarily elongating testing time. It is used to complement constructed response test because of its objectivity in scoring the responses of the examinees.

The trend of fluctuating performance in mathematics among junior secondary school students continues to attract attention of stakeholders in the education sector. Despite the effort of researchers to identify factors responsible for this below average performance and proffer possible solutions, the performance rate has still not improved. This lower performance might hinge on how the assessment practices in terms of test scoring frameworks mar examinees' performance. The wrong assessment framework adopted by the public examining body and teachers might account for one of the reasons examinees' performances in BECE mathematics were fluctuating over years. More so, assessment is one of the major tools used by global community to collect data for decision making. These assessment data refer to tests administered on a large scale or classroom settings that are designed to evaluate examinees' abilities on various concepts. Usage of assessments can be found in both developed and developing countries, although methods used differ. Assessments are used extensively in the field of education globally. In education, testing is used at many levels for instance, when a school teacher develops a classroom test items and uses learning outcomes to determine relative standing of an examinee in the class. Large scale assessments are administered at the state or national level. These tests and the decisions resulting from them can have significant impact on society at large. The procedure often used by classroom teachers and public examining bodies especially in Nigeria and other African countries, from item development up to scores generated, used traditional methods, which are fraught with many shortcomings such as sample dependent, assume equal error of measurement across the group etc. It is necessary for teachers in schools and public examining bodies to embrace modern testing theory, which is item-based for the development and scoring of examinees. The results from educational assessments are used not only to measure examinees' learning but to assess the effectiveness of teachers and schools.

More importantly, methods used to analyse examinees' responses on an assessment by the school teachers and examining bodies are technical aspects of the testing system. In educational assessment, examinees' ability is a latent trait that is not directly manifest. Instead, observable outcomes such as examinees' responses on an assessment are used to estimate the unobservable latent trait of interest. Item response theory (IRT) is an approach typically used with large scale and teacher-made assessments to model the relationship between examinees' responses and examinees' ability. IRT models can be used with various item types but in the classroom teacher-made test, they are often applied to multiple choice items while public examining bodies use multiple choice and constructed response test.

In educational measurements, two frameworks are used through which valid and reliable scores can be achieved and used for assessing examinees' performance. These are classical test theory [CTT] scoring and item response theory [IRT] scoring. Under the CTT, the examinee's test score would be the sum of the scores received on all the items in the test. This, according to Adegoke (2014), is referred to as number-correct scoring. This method of scoring produces maximum likelihood trait estimates based on raw scores (that is, total number of correctly answered item). In this method, examinees who answer correctly the same number of items irrespective of the items' level of difficulties and discriminations earn the same scale score. Thus, the nature of the items' parameters (difficulty and discrimination indices) are not considered in the scoring of examinees' performance.

IRT is a theory of testing that establishes relationship between an examinee's latent abilities and the probability of the examinee responding to a certain item correctly, estimates the parameters involved, explains the processes, and predicts the result of such an encounter (Hambleton, Swaminathan, & Rogers, 1991). More importantly, the theory is mainly interested in whether an examinee gets an item correct and not in the raw test scores, which is referred to as item-pattern scoring procedure. This scoring method produces maximum likelihood trait-estimate based on pattern of item responses (Adegoke, 2014). To calibrate test items effectively, it is important to put estimation of item parameters into consideration (that is item discrimination, denoted as 'a'; item difficulty, denoted as 'b'; and guessing or chance factor, denoted as 'c'). Consequently, the value of item parameters and ability depends on the choice of parameter model (Baker, 2001). Item parameter estimate through IRT is invariance of the features of both examinees to which it is exposed and other items that constitute the test. There are three foremost IRT applications for modelling the test data: 1PL, 2PL, and 3PL (parameter logistic) models.

While there is only one parameter ascribed to the trait level of the individual, the task or item is often characterized by the three parameters. The individual trait level is often designated by theta (θ), which represents the amount of ability, trait, or attribute level possessed by an individual. The three parameters associated with the item are (a) discrimination power, (b) the difficulty parameter, and (c) the guessing parameter (Nenty, 2000). In a cognitive task, the a-parameter indicates the degree to which examinees' response to an item varies with or relates to their trait level or ability. The b-parameter is the amount of trait inherent in an item. It represents the cognitive resistance of the item or task. In other words, it is the amount of trait under measurement just necessary to overcome the task or item. The c-parameter is the probability that an examinee possessed low trait in responding to an item correctly (Nenty, 2000).

Perusal of literature showed that in Sub-Saharan African countries, such as Nigeria, limited their testing within the confine of classical test theory application for the development, item analysis, and scoring of individual examinees (Adedoyin, 2010; Adegoke, 2014; Umobong & Jacob, 2016). However, in other areas of the world, such as the United Kingdom (UK), Ireland, United States of America (USA), and Germany, IRT application for large scale testing and scoring procedure has witnessed tremendous acceptance in measurement and research practice (Courville, 2004; Fan, 1998; Fitzpatrick & Yen, 1995; Stage, 2003; Yen & Candell, 1991). CTT and IRT theoretically connote two different contrasting frameworks; therefore, it is expected that using just any of the two frameworks without assessing assumptions and appropriate scoring framework will surely affect the final scores of the examinees. For instance, if CTT method of scoring is used, which looked at the ability of examinee based on the total score rather than looking at examine ability based on each of item, it could lead to disparity and inaccuracy in the scores of examinees coming from CTT rather than IRT frameworks. Several empirical studies concluded that there existed a statistically significant difference in the test scores of the examinees using the two contrasting frameworks (Fitzpatrick & Yen, 1995; Wilberg, 2004; Yen & Candell, 1991). However, the position of Adegoke (2014) and Courville (2004) disagreed with the earlier submission. They found out that there was no statistically significant difference in the examinees' overall mean test scores under CTT and IRT scoring frameworks. In these studies, examinees' ability was established using different test data such as ACT assessment test, public examining external tests, and teacher-made tests. Despite their propositions on comparability of examinees' ability estimate under CTT and IRT frameworks, none of their studies considered investigating ability estimate of *Basic Education Certificate Examination* of junior secondary school 3. Hence, it is imperative to carry out a study in this area.

Research Questions

Three research questions were advanced for this study.

1. Do test data of 2017 BECE mathematics fulfils dimensionality and item local independence assumptions of IRT method?
2. What are the estimated examinees' mean test scores of 2017 NECO BECE

mathematics using the two contrasting frameworks (CTT and IRT)?

3. Is there any significant difference in the test scores of the examinees using the two contrasting frameworks (CTT and IRT)?

Methods

This study adopted survey research design. The population for the study comprised of private junior secondary three (JSS3) students who enrolled for *Basic Education Certificate Examination* (BECE) in Osun State, Nigeria. An intact class of junior secondary school 3 was drawn randomly from Osogbo and Olorunda local government areas (LGAs) making 978 examinees altogether. Among the 978 sampled participants, 425 (43.5%) were boys, 553 (56.5%) were girls, and their mean age was 12 years. The instrument used was the 2017 NECO *Basic Education Certificate Examination Mathematics Multiple Choice Test*. The test consisted of 60 items each having five response options which were dichotomously scored as right or wrong. Data were analysed three ways: (a) expected a posteriori (EAP) for the establishment of person scoring used jMetrik™ software (Psychomeasurement Systems, 2018); (b) non-linear factor analysis used normal ogive harmonic analysis robust method (NOHARM) software version 4.0 (McDonald, 1997, 1999); and (c) conditional independence was assessed using Yen Q3 statistics implemented in jMetrik™ software.

Results

Dimensionality and Item Local Independence Assumptions of IRT Method

Assessment of unidimensionality assumption of 2017 National Examinations Council BECE mathematics items was done using non-linear factor analysis implemented in normal ogive harmonic analysis robust method (NOHARM statistical software). NOHARM is a computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait based on theory developed by McDonald (1997, 1999). Similarly, its ability to determine the actual number of factors embedded in the test data through model-data fit indices cannot be overemphasized. More importantly, from the result of the analysis, there is what we called residual matrix (lower off-diagonal), which is the point where NOHARM model-data fit information are found. It produces this residual matrix to aid model-data fit analysis. The residual matrix establishes the difference between the observed covariances and that of the items after the model has been fitted to the data. Thus, the best condition is where the differences are zero (0).

In Table 1 below, it can be observed that unidimensional solution’s residuals are relatively small compared to the item covariances. More so, scrutiny of the residual matrix does not disclose any large residuals. Therefore, to review the residual matrix, NOHARM provided its root mean square (RMS). The RMS is the square root of the average squared difference between the observed and predicted covariances. Thus, root mean square with small values indicates good fit. McDonald (1997) suggested that the overall measure of model-data fit may be evaluated by comparing it to four times the reciprocal of the square root of the sample size, which can be expressed mathematically as RMS criterion $= 4 \frac{1}{\sqrt{sample\ size}}$

In this study, the sample size was 978, and this size gave RMS criterion of 0.128. Thus, if the estimated value from root mean square (RMS) residual (0.022) was significantly small compared to that of RMS criterion (0.128), the conclusion is that the test data are measuring only a single construct. Another measure of dimension is Tanaka’s (1993) goodness-of-fit index

(GFI). McDonald (1999) suggested that a GFI of 0.90 indicates an acceptable level of fit, a value of 0.95 indicates good fit, and GFI of 1.00 indicates perfect fit. Therefore, the estimated GFI (0.9009) indicates an acceptable level of fit. It can be observed, based on the aforementioned indices, that one-dimension model fits the data substantially. Similarly, reliability test analysis was used to corroborate the result from the model-data fit using NOHARM for establishing unidimensionality. Guttman's L2 gave reliability coefficient of 0.799, and the standard error of measurement (SEM) was 3.2134 with 95% confidence interval. These results indicated that the 2017 BECE mathematics was unidimensional.

Table 1

Residual Matrix (Lower Off-Diagonals)

Items	47	48	49	50	51	52	53	54	55
47	0.042								
48	-0.001	0.001							
49	-0.008	0.004	0.011						
50	7.2e-6	-0.001	-0.009	-0.004					
51	-0.037	-0.043	-0.002	-0.009	-0.029				
52	0.047	0.037	-0.008	0.011	0.035	0.028			
53	-0.001	-0.008	-0.002	0.007	0.002	-0.007	-0.037		
54	0.039	0.035	0.003	0.011	0.034	0.027	-0.015	-0.042	
55	0.020	0.015	-0.001	2.3e-4	0.022	0.005	-0.014	-0.061	-0.019
+									
59	-0.004	-0.002	-0.007	-0.031					
60	-0.008	0.003	-0.039	0.037	0.032				

Notes: Sum of squares of residuals (lower off-diagonals) = 0.8199; Root mean square of residuals (lower off-diagonals) = 0.0215; Tanaka index of goodness-of-fit = 0.9009

Yen Q_3 statistics was used to establish item local independence of test data implemented on jMetrik™ software. Yen Q_3 statistics is the correlation of residuals for a pair of items after the person location estimates are controlled for. After obtaining the residuals, the linear correlation between the residuals from pair of items (say item 1 and 2, item 1 and 3, item 1 and 4 and so on until all the items in the test are all paired) is then examined to find pairs of items with large residual correlations. In this study, 1770 linear correlations were established using jMetrik™ software. Correlation coefficient larger than 0.2 screening criterion suggested by Yen (1993) indicated that the paired item violates local item independence. Therefore, it can be observed in Table 2 below that Q_3^s show that about twenty-five (25) item pairs (such as item 2 and 4, item 56 and 59, item 56 and 60 and so on) have absolute value exceeding the screening value of 0.2. This result implies that, after fitting the unidimensional 3PL model to the data, the items in these pairs had slightly more than 10% of their residual variability in common. Thus, these item pairs were considered to be exhibiting item dependence and evidence of conditional dependence in the remaining 1745 pairs was absent. These results gave evidence that the test data meet conditional item independence. Table 2 presents Q_3 statistics (edited) among the 60 items contained in the 2017 NECO BECE mathematics multiple choice exam.

Table 2
Q_3 Statistics for the NECO BECE Mathematics Test Items

Items	1	2	3	4	5	56	57	58	59	60
1	1.0000									
2	-0.0123	1.0000								
3	-0.2013	0.2203	1.0000							
4	-0.1560	0.3863	0.1792	1.0000						
5	-0.0817	-0.0866	0.1754	-0.1731	1.0000					
+										
56	-0.0235	0.0524	0.0148	0.0120	0.0644	1.0000				
57	0.0404	0.0633	0.0874	-0.1321	0.0187	-0.2492	1.0000			
58	0.0408	0.0512	0.1041	0.0061	0.0015	0.2111	0.1936	1.0000		
59	-0.1130	0.0128	-0.0548	-0.0115	0.0861	0.4147	0.0040	0.0648	1.0000	
60	-0.0736	0.0136	0.0313	0.0287	0.0764	0.3382	-0.1746	0.0187	0.0465	1.0000

Estimated Mean Scores Using CTT and IRT Frameworks

The examinees' scores in the 2017 NECO BECE 60-mathematics test items were examined, and the raw scores in the CTT model and IRT model were converted to scale score using z-score and t-score respectively. Similarly, the examinees' test score under the CTT framework (number correct scoring) and IRT framework (item pattern scoring) were both converted to the same metric using t-score (t = 10z + 50). The examinees' test score under CTT framework was first transformed to z-score, using the equation $z = \frac{X - \mu}{\sigma}$, where x = examinee's test score, μ = mean of the test scores obtained by all examinees, and σ = standard deviation of the test scores obtained by all examinees. After that, the scores were transformed to t-score. The examinees' ability estimate (in z-score) under IRT was transformed to t-scores. Then the overall mean test scores were obtained. Table 3 presents the mean, standard deviation (SD), and minimum and maximum scores of the examinees' scores in the 2017 NECO BECE 60-mathematics test item under CTT and IRT scoring frameworks.

It can be observed from Table 3 that the examinees' mean score and standard deviation under number-correct scoring method was 50.70 (SD = 10.30) and 47.78 (SD = 8.49) under item-pattern scoring. The mean difference was 2.29. Also, paired-samples *t*-test statistics showed that the mean difference was not statistically significant (t = 2.635, df = 977, P = 0.01).

Table 3
Descriptive Statistics of the Examinees' Scores in the 2017 NECO BECE Mathematics

Statistics	Number-correct scoring			Item-pattern scoring	
	Raw score	Z-score	T-score	Z-score	T-score
Minimum	5.00	-2.05	29.51	-0.46	45.40
Maximum	33.00	2.63	76.35	3.78	87.68
Mean	17.25	0.07	50.70	-0.21	47.78
SD	5.98	1.00	10.30	0.85	8.49

Significant Differences Using the Two Frameworks

The scores of six examinees who had a raw score of 31 were examined. Table 4 presents the item parameters (discrimination and difficulty indices) as well as the pattern of responses of the students to the BECE 60 mathematics test items. The table also shows examinees' corresponding number-correct (NC) and item-pattern (IP) scores in terms of z-score and t-score.

Table 4
Items Parameters

Item Number	Item Parameter		Six examinees with raw score (NC) of 31					
	b	*a*	*B1*	*B2*	*B3*	*B4*	*B5*	*B6*
1	2.03	0.82	0	0	1	1	1	1
2	1.67	2.86	0	0	0	0	0	0
3	0.70	2.85	1	1	1	1	1	1
+	+	+	+	+	+	+	+	+
58	5.91	1.23	1	1	1	1	1	1
59	1.77	0.82	1	1	0	0	0	0
60	8.50	0.30	1	0	1	1	0	0
NC Score			31	31	31	31	31	31
Z-score (IP)			0.9530	-1.0535	1.2622	2.5366	1.1407	1.8782
Z-score (NC)			2.3003	2.3003	2.3003	2.3003	2.3003	2.3003
T-score (NC)			73.003	73.003	73.003	73.003	73.003	73.003
T-score (IP)			59.530	39.465	62.622	75.366	61.407	68.782

It can be observed in Table 4 that each of the examinees had a raw score of 31 when the number-correct scoring method was used. On conversion to z-score and t-score, all the examined examinees had the same scores, 2.3003 and 73.003 respectively. However, when the item-pattern scoring method was used in estimating the examinees scores, significant variations in the scores emerged. Candidates B1, B2, B3, B4, B5, and B6 had z-score (person score, obtained from jMetrik™ software) of 1.1407, 1.8782, 0.9530, -1.0535, 1.2622, and 2.5366, respectively. When these scores were converted to t-score, candidate B6 had the highest score with 75.366, followed by candidate B2 with score of 68.782, followed by candidate B5 with score of 62.622, followed by candidate B1 with score of 61.407, followed by candidate B3 with score of 59.530, while candidate B4 had the lowest scores of 39.465.

Discussion

The significance of IRT assumptions in educational measurement is inevitable. These assumptions need to be assessed before any further analysis could be carried out on test data. These issues include dimensionality and item local independence. The choice of which IRT model to be used when calibrating test under IRT measurement framework is determined by the number of dimensions embedded in the test data. Thus, the results suggested that test data of NECO BECE 2017 mathematics items satisfy dimensionality and item local independence assumptions of item response theory. Also, it was found that examinees' mean scores' difference between classical test theory and item response theory was statistically significant. This finding is in line with other researchers (Adedoyin, 2010; Fitzpatrick & Yen, 1995; Yen & Candell, 2011) who found statistically significant differences in the examinees' mean test scores using CTT and IRT scoring approaches. Findings of this study disagree with the findings of Adegoke (2014) and Courville (2004) that no difference observed in the mean scores using CTT and IRT methods. The results also showed that IRT method of scoring produced different test scores for candidates who have the same raw scores under the classical test theory method. The differences observed in the test scores of the candidates under the IRT method of scoring emanates from the disparity in the discrimination and difficulty indices of the 2017 NECO BECE mathematics items. Examinees answered different items of the test correctly. This result is because item statistics are always taken into consideration in the process of estimating examinees' test scores under IRT scoring method. This finding gives credence to the findings of Adedoyin (2010). The researcher found that IRT method of scoring produced different test scores for examinees that had the same raw score under classical test theory method of scoring.

Conclusion and Recommendation

The scoring method adopted by National Examinations Council (NECO) had been the classical test approach despite its ineptitude to estimate correctly the actual ability of the examinees. This method of scoring is neither valid and nor reliable because examinees attempted different sets of items with different psychometric properties (that is difficulty and discrimination indices). Consequently, the study concluded that CTT and IRT examinees' mean score were not comparable, and item response theory method of scoring produced different test scores for examinees that had the same raw score under classical test theory method of scoring. Therefore, it can be recommended that post-primary school teachers, NECO, and other countries that are still operating within the confine of traditional scoring method should shift their paradigm to modern method of scoring, which takes into consideration item parameters indices during estimation of examinees test score.

References

Adedoyin, O. O. (2010). Investigating the invariance of persons parameter estimates based on classical test and item response theories. *International Journal on Educational Sciences, 2*(2), 107–113. doi: 10.1080/09751122.2010.11889987

Adegoke, B. A. (2014). Effects of item-pattern scoring method on senior secondary school students' ability scores in physics achievement test. *West African Journal of Education, XXIV*, 181–190.

Baker, F. B. (2001). *The basics of item response theory*. Retrieved from http://echo.edres.org:8080/irt/baker/final.pdf

Courville, T. R. (2004). *An empirical comparison of item response theory and classical test theory item and person statistics.* (Unpublished doctoral dissertation). Texas A & M University, College Station, Texas.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. doi: 10.1177/0013164498058003001

Fitzpatrick, A. R., & Yen, W. M. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement, 32*(3), 243–259. doi: 10.1111/j.1745-3984.1995.tb00465.x

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (Eds.). (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Psychomeasurement Systems. (2018). jMetrik™ software. Retrieved from https://itemanalysis.com/jMetrik™-download/

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.). *Handbook of modern item response theory* (pp. 258–270). New York: Springer.

MacDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Nenty, H. J. (2000). Some factors that influence students' pattern of responses to mathematics examination items. *BOLESWA Journal of Educational Assessment, 17*, 47–58.

Stage, C. (2003). *Classical test theory and item response theory: The Swedish experience.* Retrieved from https://www.umu.se/globalassets/organisation/fakulteter/samfak/institutionen-for-tillampad-utbildningsvetenskap/hogskoleprovet/publications/60581_em-no-42.pdf

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage

Umobong, M. E., & Jacob, S. S. (2016). A comparison of classical and item response theory person/item parameters of physics achievement test for technical schools. *African Journal of Theory and Practice of Educational Assessment*, *4,* 115–131

Wiberg, M. (2004). Classical test theory vs item response theory: An evaluation of the theory test in the Swedish driving-license test. *Educational Measurement Working paper:* EM No 50, UMEA University.

Yen, W.M. & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five scoring metrics. *Measurement in Education, 4*(3), 209–228. doi: 10.1207/s15324818ame0403_2

Authors

Musa Adekunle Ayanwale is a Ph.D. student in the Institute of Education, University of Ibadan, Nigeria. He holds a B.Sc. (statistics) and M.Ed. (educational evaluation). He is interested in educational measurement theories and structural modelling. Also, he is an experienced quantitative analyst and distinguished author who has published in many reputable local and foreign journals. Ayanwale can be contacted at kunleayanwale@gmail.com

Joshua Oluwatoyin Adeleke is a senior research fellow in the Institute of Education, University of Ibadan, Nigeria. He holds a BSc.Ed (mathematics and statistics), M.Ed. (guidance and counselling), and Ph.D. (educational evaluation). He is an experienced mathematics educator, quantitative analyst, and distinguished author who has published in many reputable local and foreign journals. Adeleke can be contacted at jo.adeleke@ui.edu.ng

Titilayo Iyabode Mamadelo is a Ph.D. student in the Institute of Education, University of Ibadan, Nigeria. She holds a B.Ed. (adult education & English) and M.Ed. (educational evaluation). She works at National Examinations Council (NECO). Mamadelo can be contacted at titimamadelo@gmail.com