EDUCATUM JSMT Vol. 8 No.1 (2021)
ISSN 2289-7070 / e-ISSN 2462-2451 (57-69)
https://ejournal.upsi.edu.my/index.php/EJSMT/index

Calibration of Polytomous Response Mathematics Achievement Test Using Generalized Partial Credit Model of Item Response Theory

Musa Adekunle Ayanwale

Department of Education Foundations,
Faculty of Education, Kampala International University,
Kampala, 20000, Uganda

**Corresponding author: adekunle.ayanwale@kiu.ac.ug*

Published: 04 June 2021

To cite this article (APA): Ayanwale, M. A. (2021). Calibration of Polytomous Response Mathematics Achievement Test Using Generalized Partial Credit Model of Item Response Theory. *EDUCATUM Journal of Science, Mathematics and Technology*, *8*(1), 57-69. https://doi.org/10.37134/ejsmt.vol8.1.7.2021

To link to this article: https://doi.org/10.37134/ejsmt.vol8.1.7.2021

Abstract

Thisstudy assessed the empirical comparability of item calibration in the developed essay (DEV_{essay}-MAT) and ($NECO_{essay}$-MAT) mathematics achievement test under the Generalized Partial Credit Model (GPCM). The instrumentation research approach of counterbalance design was employed. The sample consisted of 1080 senior secondary school students (SSS3) of 36 schools, who were drawn randomly from Osun East senatorial district of Osun State, Nigeria. Two instruments were used and data obtained were subjected to Parallel Analysis (PA), Generalized Partial Credit Model (GPCM) and Independent sample t-test. Results showed that the test does not violate unidimensionality with the first Eigenvalue (2.05) from the experimental data was greater than the first random Eigenvalue (1.17) from PA, while other Eigenvalues from the experimental data were less than the rest of Eigenvalues under PA. Also, there existed a significant difference between the step difficulties/overall item difficulty and discrimination/slope index of the two instruments with ($t = 3.52$, $df = 8$, $p < 0.05$) and ($t = 3.26$, $df = 8$, $p < 0.05$) respectively. The author concluded that the developed essay test produced better item statistics estimates compared to NECO-MAT (essay) test. Consequently, it was recommended that public examining bodies in sub-Sahara Africa should embrace an apt polytomous model for the calibration of their test items.

Keywords:*Generalized Partial Credit Model; Polytomous test;Item Response Theory; Item Calibration; Parallel Analysis*

INTRODUCTION

Mathematics is an essential component of human logic and reasoning, as well as attempts to understand the universe and ourselves. It is a great way to develop mental discipline while encouraging logical thinking and mental rigor. Mathematics, an exemplification of knowledge, apparatus and dialect of science required to create ways of managing issues, not as it were at pre- and post-basic education but in all components of human being. It is believed that science and innovation were built on this pedestal. Furthermore, understanding the contents of other school subjects such as chemistry, physics and so on are dependent on mathematical knowledge. Mathematics, according to [1], provides the experience required to improve problem-solving skills, not only in school but in all facets of life, and a well-developed science that is important in all aspects of human endeavor. As a critical tool for understanding and applying science and technology, the discipline serves as a forerunner and harbinger of much-needed technical and, of course, national development, which has become critical in the world's developing nations[2]..

Consequently, to move with fast-changing advancement in science, technology and innovation, the learning of mathematics is very imperative in every society. The importance of mathematics to humanity is enormous, which contributed to one of the reasons for its inclusion in school educational programs as one of compulsory subjects for each child of school age. This would facilitate development of well-suited scientific abilities needed by the learners to manage any life challenges [2] . More importantly, it is clear that no other subject forms such a strong force among the various branches of science. Thus, [3] described mathematics as the core intellectual discipline of technology societies. Science knowledge remains superficial without mathematics . It means, therefore, that the position of mathematics in secondary school curriculum is essential for scientific development.

In spite of the significance of Mathematics, the performance rate of examinees in the subject at the external examination administered by the National Examinations Council (NECO) continue to fluctuate over the years. The bar chart (Figure 1) delineates investigation of examinees' performance spanning between 2005 to 2015.

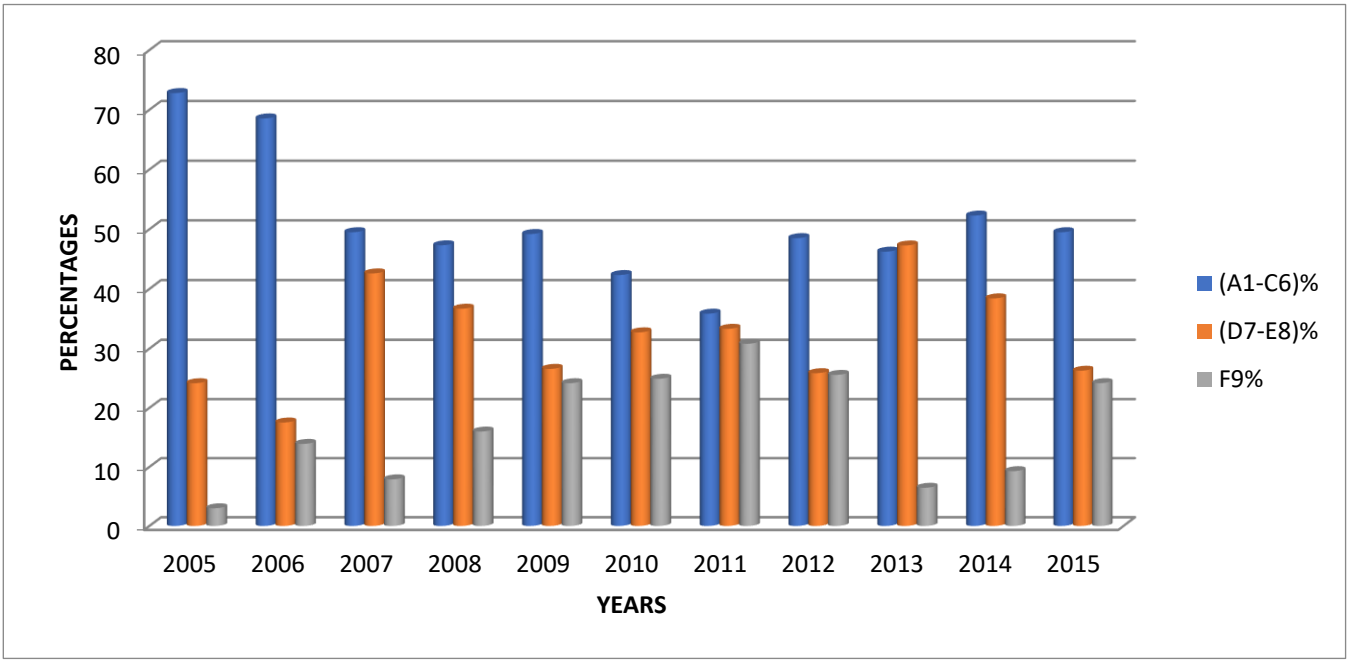

Figure 1: Examinees' performance in SSCE Mathematics administered by the National Examinations Council (NECO) between 2005 and 2015

Quick look at the performance portrays that less than 50% of examinees' that sat for the examination passed at credit level between (2007-2013 and 2015). However, in year 2005, 2006 and 2014 the percentage pass at credit level rose above 50%. The implication is that in the years other than 2005, 2006 and 2014, less than 50% of Nigerian students were able to secure admission into higher institutions of learning. In any case, various variables had been credited to ceaseless fluctuating performance in the subject by numerous researchers. Eminent among them include: learners' poor attitude towards mathematics [3], learners poor study habit and orientation [4], nature of the test items and examinees' characteristics [5] and poor strategies of instructing mathematics [6]. In any case, there's no problem without a conceivable solution. Noticeable among solutions proffered by diverse research studies include: the use of indigenous language in the teaching of mathematics [7] and improving the quality of instructional techniques [8]. Despite all the advanced way-out, the observed fluctuating performance of the examinees linger. However, investigating the quality of test items used by this examination body had not been explored in the literature to the knowledge of researcher, and this might account for the trend of performance observed over the years. Test items of high-stake test like this need to be valid and reliable. Thus, the characteristics of test items examinees' respond to and the inherent trait(s) being measured also have the capacity to determine what the performance would be.

It is imperative to state that public examining institutions in Nigeria such as National Examinations Council (NECO), West African Examinations Council (WAEC) etc adopted both multiple-choice and constructed response test (essay) items for their evaluation. For instance, NECO Mathematics is divided into two papers. This incorporate sixty (60) multiple-choice items which is famously known as paper III, whereas the constructed response test (essay) is known as paper II comprises of two sections. Section "A" comprised of compulsory five (5) test items and "B" comprised of seven (7) items out of which any five (5) items would be answered by the examinees. These two forms of evaluation are used by examining bodies to complement each other. In Nigeria, validity and reliability of different types of testing have promted a wide used of multiple-choice test items at the primary, secondary and post-secondary level of education, in spite its inherent guessing tendency related with the items. Its simplicity and accurate of marking, which is important in large scale assessment makes it more proficient and freer of grading bias. Be that as it may, certain critical educational skills and types of knowledge such as expository, analytical, creative thinking and expression are too volatile to ever be measured satisfactorily with multiple-choice item questions [9]. Subsequently, in this study, the constructed response test (essay/polytomous) aspect of the assessment is emphasised. This test are quick and easy to develop but they take longer time to answer and a great deal of time to mark, especially where there are large numbers of examinees. The scoring procedure regularly requires subjective judgment; different examiners may grant distinct scores to a similar response. This, however, can be addressed by inter-rater scoring.

More importantly, public examining bodies in sub-sahara Africa such as NECO is expected to determine the characteristics of test items used for their assessment. It is never inconceivable that part of reasons capable for continue below average performance might be associated to methods used by the examining bodies during their test development and item analysis respectively. As observed by Chief Examiners' Report [10,11] that examinees continually performed exceptionally poor in constructed response (essay) questions that demand applying numerical problems. Their reports, further remarked that many of the examinees' shown inadequate proficiencies in the use of principles in resolving difficult questions. At this point, the question is that do National Examinations Council (NECO) establish psychometrics properties of their constructed-response test (essay)? This may be another area which research in mathematics education has not truly centered in the time past. Therefore, according to [9], test items that is not well developed might affect examinees' performance adversely.

There are two modern approaches through which quality tests can be developed. These are Classical Test Theory (CTT) and Item Response Theory (IRT) measurement frameworks. The two test theories are used to measure behaviour and numerical values given to the behaviour for evaluation. Classical test theory as the premise of a testing theory is the only testing framework accessible to test developers and psychometricians for decades, but characterised with shortcomings such as item statistics is group-dependent, assume equal measurement error, and person statistics is test- dependent. However, Item response theory (IRT) framework was developed to adjust all the pitfalls related to CTT. As commented by [12], item response theory contains models which explain interaction of a person with a given trait level to the characteristics of the item constructed to stimulate the level to which individual examinee has that proficiency. Thus, IRT has the capability to develop quality constructed response test items (essay/polytomous).

Items that are scored on a multi point categories are alluded to as polytomous scored items. A polytomously scored item is the likelihood of an examinee achieving a particular score category which can be described by any of the polytomous IRT models or any situation in which partial credit might be awarded to indicate differing levels of item performance; they are less commonly used than dichotomous scored items. According to [13], polytomous items is that, since it contains more response categories on the trait's continuum, it provides better information over a broad range compare to multiple-choice items. As observed by [14, 15] that the entire purpose of using more than two categories per item is to obtain more information about the trait level of the examinees being measured, so that more accurate trait-level estimates can be obtained. In addition, [16]

pointed out that more detailed diagnostic information about respondents and items can be obtained from polytomous test items.

Polytomous tests are categorical items which are the same way as dichotomous items; they basically have more than two conceivable response categories. Categorical information can be depicted suitably in terms of the number of categories into which the information can be put [17, 13]. Ordered categories are characterised by boundaries or limits that separate the categories. Ordinarily, there's persistently one less boundary than there are categories. For instance, a dichotomous item requires because it was one category boundary to partition the two likely response categories. Furthermore, a four-point Likert type (that's strongly agreed, agree, disagree & strongly disagree) item requires three boundaries to section the four likely response categories. In this study, the developed response test item (essay) was scored over 8, possible score categories include 0,1,2,3,4,5,6, and 7. Here, there are eight categories of scores. In any case, applying the polytomous model, seven step difficulties are evident, which depicts, eight categories minus one.

Polytomous IRT models available with various derivations and parameterisation; among them are the Partial Credit Model (PCM) [16], Rating Scale Model (RSM), a Nominal Response Model (NRM), the Graded Response Model (GRM) [18] and the Generalized Partial Credit Model (GPCM) [19]. The determination of a particular model under IRT for an item calibration is guided by either choosing the model that best fits the dataset or choosing dataset that best fits the model [20]. In any case, [20] criticised the use of dataset that fits the model because of its negative contributions to the construct and content of the test being inspected. In this study, PCM and GPCM were considered. Meanwhile, GPCMis emphasised after conducting model data fit assessment, which as an attribute of difficulty (step difficulties) and discrimination parameters.

The GPCM is one of the models of IRT developed to analyse partial credit data, where examinees responses are scored 0, 1…, n, where ‘n’ is the highest score category for the item. The model expects that each of two adjoining categories (that's ‘n’ and ‘n-1’) in a polytomous score item can be seen as dichotomous categories. The likelihood of an examinee with a certain ability level reaching the score category ‘n’ instead of ‘n-1’ can be described by a dichotomous IRT model. Assume a polytomous scored item has ‘m’ score categories. Based on the GPCM, the item has one item discrimination parameter, one location parameter, and a set of ‘m-1’ step difficulties parameters. The model was in this way generalised from the dichotomous IRT models to portray the likelihood of selecting a specific score category from all the possible scores categories for an examinee [19]. The GPC model is expressed as:

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=o}^{k} D_{aj}\left(\theta - b_j + d_{j,v}\right)\right]}{\sum_{c=0}^{m_j} \exp\left[\sum_{v=0}^{c} D_{aj}\left(\theta - b_j + d_{j,v}\right)\right]} \quad \text{………………………………………….. (1)}$$

Where D, is a scaling constant set to 1.7 which approximate the typical ogive model, aj is a slope parameter, “bj” is an item location parameter, and “dj, v” is a category parameter. The slope parameter shows the degree to which categorical responses change among items as “q” level changes. With “mj” categories, as it were “mj -1” category parameters can be recognised.

Indeterminacies in the parameters of the GPCM are resolved by setting d_j, 0 = 0 and setting

$$\sum_{k=1}^{m_j-1} d_{j,k} = 0 \quad \text{……………………………………………………………. (2)}$$

Be that as it may, [19, 21] pointed out that bj - dj,k is the point on the “q” scale at which the plots of Pj,k-1(q) and Pjk(q) meet, so characterise the point on the “q” scale at which the response to item “j” has risen

to likelihood of falling in response category k -1 and falling in response category k. In this way, a theoretical item category response functions (what is known as item characteristic curves for dichotomous items) for GPCM item parameters is presented (see figure 2).

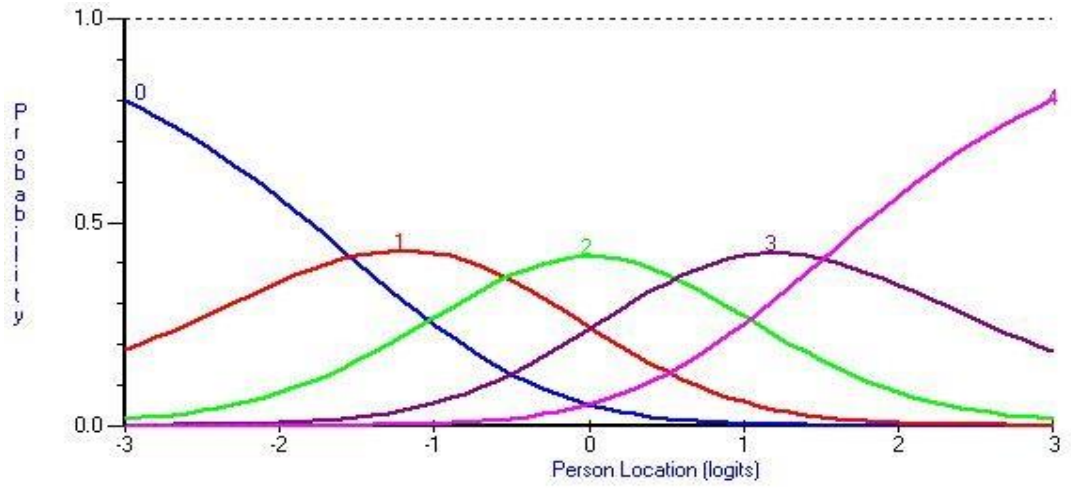

Figure 2: Hypothetical Representation of Item Category Response Functions Curve

Survey of literature has shown that in Nigeria, down to earth applications of IRT models for item calibration has been overwhelmingly utilised under multiple-choice items and public examining bodies in sub-sahara Africa such as NECO, WAEC etc were not exempted from this trend, while less attention was paid to calibration of polytomous test items, which complement the multiple-choice items used by the examination bodies for their high-stake testing. For instance, [5, 22-30] worked broadly on calibrations of multiple-choice items of different subjects. Whereas researchers [18, 19, 31-33] in developed nations like USA, UK, Germany, Ireland etc have grasped this approach of utilising IRT application for the foundation of constructed-response test items (polytomous) parameters. More importantly, the objectives of this paper are to examine the dimensionality of the test, and item facilities (steps difficulties and slope) of the tests. Based on this premise, this study explored the comparability of item calibration in the developed essay (DEV_{essay}-MAT) and $NECO_{essay}$-MATmathematics achievement test under generalised partial credit model. More so, research questions advanced to guide the study were in threefold: Do constructed mathematics achievement test items fulfill dimensionality assumption of Item Response Theory? Is there any significant difference between the step difficulties in the DEV_{essay}-MAT and $NECO_{essay}$-MAT using GPCM? And is there any significant difference between the discriminating index in the DEV_{essay}-MAT and $NECO_{essay}$-MAT using GPCM?

MATERIALS AND METHODS

Design, Population and Sample

Counterbalance design of instrumentation research type was used. The population comprised of mathematics examinees in Senior Secondary School III (SSS3) in all schools that had presented examinees for National Examinations Council (NECO) examination in the last five years in Osun State, Nigeria. Sample selection was carried out using simple random sampling method to select 6 out of 10 Local Government Areas (LGAs) from Osun east senatorial district of Osun State, Nigeria. Moreover, six (6) co-educational public schools were drawn in each of the chosen LGAs, totaling thirty-six schools, from which an intact science class was used. In this way, one hundred and eighty (1080) SSS3 examinees participated in the study. Their ages ranged between 16 and 20 years with 655 (60.6%) boys and 425 (39.4%) girls respectively.

Instrument

The initial draft of mathematics constructed response test (essay) consisted of twenty (20) test items which were constructed by the researcher. The researcher adopted senior secondary school mathematics curriculum for the construction of essay achievement test similar to National Examinations Council (NECO) essay questions. The test items covered all the topics in the mathematics curriculum for the senior secondary school examination for the Nigerian students (see appendix; for detailed topics under each of the theme). These are: Theme 1 -number and numeration, theme 2-algebraic process, theme 3- geometry, theme 4 -statistics and probability. The drafted constructed response tests were subjected to experts review who were examiners of National Examinations Council as well as experienced secondary school mathematics teachers for their vetting. Corrections in terms of ambiguity and clarity of words were strictly adhered to re-write the test. Table of specification for the draft fifteen constructed response test (essay) consisting of twenty-six (26) sub-independent items is presented (see Table 1).

Table 1: Fifteen-Draft Essay Mathematics Achievement Test (DFT-MAT_{essay})

Content	Behavioural Objectives			Total
	Knowledge (0%)	Comprehension (0%)	Higher Order (100%)	
Number& Numeration (23%)	-	-	6 (Items 2a, 3a, 3b, 7a, 11b, 12)	6
Algebraic Process (42%)	-	-	11 (Items 1, 2b, 6a, 6b, 7b, 10b, 11a, 13b, 14a, 14b, 15)	11
Geometry (23%)	-	-	6 (Items 5, 8a, 8b, 9a, 9b, 13a)	6
Statistics& Probability (12%)	-	-	3 (Items 4a, 4b, 10a)	3
Total	-	-	26	26

Thereafter, these items were administered to population of examinees of senior secondary school 3 who were outside the target population for this study. Item analysis of item response theory was conducted on the data obtained for the fifteen essay items and five (5) best surviving test items formed the final developed essay items (DEV_{essay}-MAT). Consequently, two instruments were used for data collection: Self-developed Mathematics essay (DEV_{essay}-MAT) with content validity index of 0.79 and empirical reliability of 0.85, and $NECO_{essay}$-MAT with content validity index of 0.60 and empirical reliability of 0.71 respectively. Also, the marking rubric provides the examiner what features of the response to focus and how to determine how many points to award to a response. This scoring rubrics was developed by 10 experienced mathematics examiners, while the validity and reliability of the rubrics was established using lawshe content validity index with 0.83 and inter rater relibility coefficient of 0.81. Each of the question was scored over 8, with possible score categories of0,1,2,3,4,5,6,7, and 8. In essence, there are nine categories of scores. Using polytomous model, 8 step difficulties were evident, which depicts, nine categories minus one.

Data Collection

Data collection was carried out using counterbalance under the same setting. That is, in each of the schools that was selected, the population was divided into two groups (group I and II) and two categories of test forms (Test A and B). The first possible order was when group I were responding to Test A, group II were responding to Test B. The second possible order of administration was to present group I with Test B and group II with Test A at the same time. This approach enabled the researcher to measure the effects in all possible

situations and boost the validity of data for the study. Table 2 presented schematic diagram of counterbalance design

Table 2: Schematic Diagram of Counterbalance Design

Population	Sample	X_1	X_2	Y_1	Y_2
P	1	✓			✓
P	2		✓	✓	

Data Analysis

Obtained data was analysed using Parallel Analysis (PA) implemented in Monte Carlo PCA software version 2.3.0 for the establishment of dimensionality, Generalized Partial Credit Model (GPCM) of item response theory was used for calibration implemented in IRT-PRO Version 4.0.1 while independent sample t-test at 0.05 significant level was used to compare the means difference across the item parameters (that is step difficulties and discriminating index).

RESULTS

Findings on dimensionality

Assumption of dimensionality of IRT utilizing parallel analysis was conducted through Monte Carlo Principal Component Analysis for parallel analysis computer program. This requires that a set of arbitrary relationship matrices be created based upon the same number of factors and respondents as the experimental data. These random correlation matrices are at that point subjected to principal components investigation and the average of their eigenvalues are computed and compared to the eigenvalues created by the experimental data [34]. Table 3 presented the Monte Carlo PCA for parallel analysis statistic.

Table 3: Monte Carlo PCA for Parallel Analysis Statistic for DEV_{essay}-MAT

Components	Experimental Eigenvalues	Random Eigenvalues	Standard Deviation
1	2.052	1.175	0.021
2	1.119	1.153	0.016
3	1.115	1.137	0.013
4	1.098	1.115	0.012
5	1.019	1.062	0.011
6	0.984	1.038	0.010
7	0.945	1.018	0.010
8	0.933	0.998	0.009
9	0.898	0.977	0.009

10	0.882	0.957	0.011
11	0.831	0.936	0.010
12	0.799	0.915	0.012
13	0.776	0.892	0.012
14	0.747	0.868	0.013
15	0.702	0.836	0.016

As observed in Table 3, the first component eigenvalue (1.175) was not greater than the first eigenvalue of the experimental data (2.052) while other eigenvalues from the experimental data were less than the second, third, fourth and fifth eigenvalues. This suggests that DEV_{essay}-MAT items were unidimensional. This laid credence to [34] recommended criteria for assessing unidimensionality. Consequently, further analysis can be conducted on the data.

Findings on item calibration (Difficulty index)

The DEV_{essay}-MAT and NECO-MAT (essay) were calibrated with Generalized Partial Credit Model using Marginal Maximum Likelihood estimation implemented in Multivariate EQSIRT software version 2.1. The overall mean difficulty and step difficulties for each item were estimated. Each essay item was scored over 8, with score categories of 0,1,2,3,4,5,6, and 7. Therefore, since there were eight (8) categories of score, there would be seven (7) step difficulties (b1, b2, b3……., b7). Tables 4, 5 and 6 present the item parameters (that is discrimination and step difficulties) and independent sample t-test statistic of DEV_{essay}-MAT and NECO-MAT (essay) under IRT framework.

Table 4: Item Parameter Statistics of DEV_{essay}-MAT

Items	***a (slope)***	***Location***	b_1	b_2	b_3	b_4	b_5	b_6	b_7
1	0.93	0.23	-1.52	-1.43	0.39	0.78	0.82	0.99	1.98
2	0.89	-0.92	-2.70	-1.41	-0.94	0.26	0.72	0.79	1.69
3	0.68	0.65	-1.86	-1.34	0.18	0.71	0.87	2.20	3.00
4	0.54	0.51	-2.45	-1.25	0.14	1.30	1.47	2.13	2.38
5	0.47	0.06	-1.79	-0.29	0.12	0.13	0.49	1.93	2.11
Mean	**0.70**	**0.11**							
SD	**0.20**	**0.62**							

Table 5: Item Parameter Statistic of NECO-MAT (Essay)

Items	**a (slope)**	**Location**	b_1	b_2	b_3	b_4	b_5	b_6	b_7
1	0.30	1.55	-1.50	0.93	1.20	1.53	2.38	2.64	2.95
2	0.14	1.50	-0.89	1.15	1.22	1.72	1.86	1.99	2.07
3	0.44	1.12	-0.47	-0.13	0.57	1.49	1.72	1.85	2.45
4	0.26	0.90	-1.65	0.32	1.05	1.16	1.24	1.42	3.00

5	0.52	0.91	-1.73	0.39	0.75	1.16	1.18	1.33	2.53
Mean	**0.33**	**1.20**							
SD	**0.15**	**0.31**							

Table 6: Independent Sample t-test statistics of Step Difficulties in the DEV_{essay}-MAT and NECO-MAT (Essay)

			Equal variances assumed step_difficulty	**Equal variances not assumed step_difficulty**
Levene's Test for Equality of Variances	F		**0.84**	
	Sig.		**0.39**	
t-test for Equality of Means	t		**3.52**	**3.52**
	df		**8**	**5.93**
	Sig. (2-tailed)		**0.01**	**0.01**
	Mean Difference		**1.09**	**1.09**
	Std. Error Difference		**0.31**	**0.31**
	95% Confidence Interval of the Difference	Lower	**0.38**	**0.33**
		Upper	**1.80**	**1.85**

Cursory look at Tables 4 and 5 depict the estimated item parameters of the developed constructed response items (essay) and the NECO-MAT (essay) test items respectively. The item location is the average estimates of the step difficulties and indicates the overall difficulty of the item. The step difficulties (b-parameters) show the point on the metric scale at which adjacent responses are equally likely. As seen in the Tables, the mean and standard deviation values for overall item difficulty of the DEV_{essay}-MAT and NECO-MAT (essay) test were ($M = 0.11$; $SD = 0.62$) and ($M = 1.20$; $SD = 0.31$) respectively. Meanwhile, as observed by [31] that like ability (θ), step difficulties have a theoretical range from $-\infty$ to $+\infty$ but are practically falls within the range of -2 and +2. This will enable the test items not to be extremely simple or extremely difficult for the projected test population. Based on this, these statistics depict that on the average, items of the NECO-MAT (essay) test were more difficult than the items of the developed constructed response test (DEV_{essay}-MAT). Also, as indicated in Table 6, the independent sample t-test statistics conducted showed that there existed a statistically significant difference with ($t = 3.52$, $df = 8$, $p < 0.05$). The implies that the two instruments were not related in terms of difficulty of the test items.

Findings on item calibration (Discrimination index)

Table 7: Independent Sample t-test statistics of Discriminating index in the DEV_{essay}-MAT and NECO-MAT (Essay)

		Discrimination Equal variances assumed	**Equal variances not assumed**
Levene's Test for Equality of Variances	F	**0.93**	
	Sig.	**0.36**	
t-test for Equality of Means	t	**3.26**	**3.26**
	df	**8**	**7.33**

Sig. (2-tailed)		**0.01**	**0.01**
Mean Difference		**0.37**	**0.37**
Std. Error Difference		**0.11**	**0.11**
95% Confidence Interval of the Difference	Lower	**0.11**	**0.10**
	Upper	**0.63**	**0.64**

The second column of Tables 4 and 5 will be referred to. These column depicts the discrimination parameter or slope (a) of the developed constructed response test (essay) and NECO-MAT (essay). This indicates how steeply the probability of categories of correct response changes as the ability increases. As seen in Tables 4 and 5, the mean and standard deviation values for the slope in the developed DEV_{essay}-MAT and NECO-MAT (essay) test were (M = 0.70; SD = 0.20) and (M = 0.33; SD = 0.15) respectively. [31] remarked that the theoretical range for slope is from $-\infty$ to $+\infty$, but the practical range is from 0 to perhaps 2 or 3. Based on this premise, the developed test item (DEV_{essay}-MAT) differentiates better between examinees with different levels of construct compare to NECO-MAT (essay) test items. More importantly, as observed in Table 7, the independent sample t-test statistics conducted showed that there existed a statistically significant difference with ($t = 3.26$, $df = 8$, $p < 0.05$). The implies that the two instruments were not at par in terms of distinguishing between examinees who know the material tested and those who do not.

DISCUSSION AND CONCLUSION

Assessment of dimensionality for either dichotomous or polytomous items within the confine of Item response theory measurement framework is an indispensable aspect of item analysis. Unidimensionality assumption, conducted on constructed response Mathematics achievement test revealed that items of the test tap into only one dimension. More so, only a single score is reported for the test items, which depicts there is an implicit assumption that the items share a common primary construct. The authors' concluded that the DEV_{essay}-MAT fulfilled the assumption of unidimensionality as the Monte Carlo PCA for parallel analysis results was in line with the set criteria for assessing unidimensionality by [34, 35] that the factor extraction is where the eigenvalues generated by experimental data exceed the eigenvalues produced by the random data. The results also showed that there existed a statistically significant difference between the developed constructed response test (essay) and test of NECO-MAT (essay) item parameters (step difficulties and discriminating index). Findings of this study corroborated studies by [19, 21, 31-33] that modelling real data with Generalized partial credit model yielded better estimates compare to simulated data. Conversely, findings from this study negates findings by [32] that GPCM produced better item parameters estimate for simulated data. The observed difference in the two studies were not far fetch. In [32] study, the dimensionality of the data used was not established and incorrect choice of the model might lead to spurious estimations of item parameters. The impact of this study is that psychometrics properties of the developed constructed response test (essay) were established and verified compare to conventional ways of moderating essay items of National Examinations Council (NECO) by using subject expert decision which is psychometrically not good. Consequently, the author of this study concluded that the NECO-MAT (essay) test was slightly difficult compare to develop constructed response test (essay) under Generalized partial credit model. Also, developed test items discriminated better among examinees with different levels of abilities than the NECO-MAT (essay) test items. Therefore, public examining bodies in sub-sahara Africa should always endeavour to establish psychometrics properties of their polytomous test items using appropriate item response theory measurement model. However, failure to do the necessary, might affect the performance of examinees and the award certificate may be questioned.

Acknowledgements

The author expresses his profound gratitude to the Director of Schools, Ministry of Education, Osun State, Nigeria for given me an enabling environment to operate in their various schools during collection of data for this study. Also, the author acknowledge that this paper was presented at the 5th International Conference of Educational Assessment and Research Network in Africa held between 1-6 September, 2019.

Declaration of Conflicting Interests

The author declare no conflict of interest.

REFERENCES

[1] Oladipupo-Abodunwa, T.O., Adeleke, J. O. & Ayanwale, M..A. (2019). Student Mathematics Engagement: Development and Validation of a Measurement Instrument, African J. Behav. Scale Dev. Res., vol. 1, no. 2, pp. 17–23.

[2] Ayanwale, M.A. & Adeleke, J.O. (2020). Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous Response Mathematics Achievement Test. Bulg. J. Sci. Educ.Policy, vol. 14, no. 2, pp. 260–285. Accessed: May 31, 2021. Available: https://www.academia.edu/45182779/

[3] Akinsola, M. K. (1994). Comparative effects of mastery learning and enhanced mastery Learning strategies on Learners' Achievement and Self-concept Mathematics. Unpublished PhD Thesis. Faculty of Education. University of Ibadan. xvii+205pp.

[4] Umameh, M. A. (2011). A Survey of Factors Responsible for Learners' Poor Performance in Mathematics in Senior Secondary School Certificate Examination (SSCE) in Idah Local Government Area of Kogi State, Nigeria. Unpublished M.Ed Dissertation. Faculty of Education. University of Ibadan.

[5] 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. *European Scientific Journal* 12:263-284.

[6] Asikhia, O. A. (2010). Learners and teachers' perception of the causes of poor academic Performance in Ogun state secondary schools: Implications for counseling for National development. *European Journal of Social sciences* 13.2: 28-36.

[7] Adegoke, B. A. (2011). Effect of direct Teacher influence on dependent-prone Learners' Learning outcomes in secondary school mathematics. *Electronic Journal of Research in Educational Psychology* 9: 283 – 308.

[8] Abina, D. B. (2014). Influence of teacher characteristics, availability and utilization of instructional materials on learners' performance in mathematics. Unpublished PhD Thesis. Faculty of Education. University of Ibadan. xiv+193pp.

[9] Adewale, J.G., Adegoke, B.A., Adeleke, J.O. & Metibemu, M.A. (2017). A Training Manual onItem Response Theory, 1st ed. Ibadan: Institute of Education, University of Ibadan in Collaboration with National Examinations Council, Minna, Niger State.

[10] National Examinations Council (2012). Chief Examiners Report in Mathematics. Retrieved on August 6, 2019 from http://www.mynecoexams.com/examiners report.html

[11] National Examinations Council (2013): Chief Examiners Report in Mathematics. Retrieved on August 6, 2019 from http://www.mynecoexams.com/examiners report.html

[12] Rupp, A. A. (2009). Item Response Theory modeling with Bilog-MG and Multilog for windows. *International Journal of Testing* 3.4: 365-384.

[13] Ostini, R.& Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks: Sage Publication.

[14] Bejar, I.I. (1997). An application of the continuous response level model to personality Measurement. *Applied Psychological Measurement* 1: 509-521.

[15] Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Psychological Measurementin Education* 1: 279-297.

[16] Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometric*,47:149-174.

[17] Ayanwale, M.A. (2019). Efficacy of Item Response Theory in the Validation and Score Ranking of Dichotomous and Polytomous Response Mathematics Achievement Tests in Osun State, Nigeria. doi: 10.13140/RG.2.2.17461.22247.

[18] Samejima, F. (1969). Estimation of Latent Ability using a Response Pattern of Graded Scores. *Psychometrica*, Monograph Supplements No. 17.
[19] Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm. *Journal of Applied Psychological Measurement* 16: 159-176.
[20] Yen, W.M. (1992). Item response theory. *Encyclopedia of Educational Research*. 6th ed. NY: Macmillian. 657-667.
[21] Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Journal of Applied Psychological Measurement* 14: 59-71.
[22] Adedoyin, C. (2010). Investigating the Invariance of Person Parameter Estimates based on Classical Test and Item Response Theories. *An International Journal on Education Science* 2: 107- 113.
[23] Adegoke (2013). Comparison of item statistics of physics achievement test using Classicaltest theory and item response theory frameworks. *Journal of Education and Practice* 4.22: 87 – 96.
[24] Adegoke (2014). Effects of Item-pattern scoring method on Senior Secondary School Learners Ability Scores in Physics Achievement Test. *West African Journal of Education* Vol. XXIV: 181-190.
[25] Ayanwale, M.A., Adeleke, J.O. & Mamadelo, T.I. (2018). An assessment of item statistics estimates of Basic Education Certificate Examination through Classical Test Theory and Item Response Theory approach. International Journal of Educational Research Review, 3(4), 55-67. Doi: 10.24331/ijere.452555.
[26] Enu, V.O. (2015). Using item response theory for the validation and calibration of mathematicsand eography items of Joint Command Schools Promotion Examination in Nigeria. Unpublished Doctoral Thesis. Institute of Education. University of Ibadan.
[27] Fakayode, O. (2018). Comparing CTT and IRT measurement frameworks in the estimation of item parameters, scoring and test equating of West African Examinations Council Mathematics Objective Test for June and November, 2015. Unpublished PhD thesis. Institute of Education, University of Ibadan.
[28] Ogbebor, U.C. (2017). Construct of Mock Economics Test for Senior Secondary School Learners in Delta State, Nigeria using Classical Test and Item response theories. Unpublished PhD thesis. Institute of Education. University of Ibadan
[29] Ojerinde D. (2013). Classical Test Theory (CTT) vs Item Response Theory (IRT): An Evaluation of the Comparability of Item Analysis Results. Paper Presentation at the Institute of Education. University of Ibadan. May 23, 2013.
[30] 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*, Vol.4,115-131.
[31] DeMars, C. (2010). *Item Response Theory. Understanding statistics measurement*. City: Oxford University Press.
[32] Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologist*. Mahwah, New Jersy: Lawrence Erlbaum Associates.
[33] Grima, A. M. & Weichun, W. M. (2002). Test Scoring: Multiple-Choice and Constructed-Response Items. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
[34] Watkins, M.W. (2006). Determining Parallel Analysis Criteria. *Journal of Modern Applied Statistical Methods*, 5(2), 344-346
[35] Ledesma, R.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. *Journal of Practical Assessment, Research and Evaluation*, *12*(2), 1-11.

APPENDIX

THEME I – NUMBER AND NUMERATION

(a) Number base system
(b) Modular arithmetic
(c) Logarithms
(d) Sets theory
(e) Sequence and series
(f) Quadratic par
(g) Simultaneous linear par
(h) Surds
(i) Matrices and determinants
(j) Arithmetic of finance

THEME II – ALGEBRAIC PROCESS

(a) Simple pars and variation
(b) Logical reasoning
(c) Gradient of a curve
(d) Linear inequalities
(e) Algebraic fractions
(f) Application of linear and quadratic pars to capital market

THEME III – GEOMETRY

(a) Constructions
(b) Proofs of some basic theorems
(c) Trigonometric ratios and trigonometry graphs
(d) Mensuration
(e) Chord property
(f) Bearings
(g) Surface zone and mass of circle
(h) Longitude and scope
(i) Coordinates geometry of straight lines

THEME IV – STATISTICS AND PROBABILITY

(a) Data presentation
(b) Measures of focal inclination
(c) Measures of scattering
(d) Histograms of grouped data
(e) Cumulative frequency graph
(f) Measures of central tendency for grouped data
(g) Probability