UNIJERPS
Unizik Journal of Educational Research and Policy Studies VOL.19 (3);
January-June, 2025 https://unijerps.org
2023 Impact Factor:* 6.834;** 2022 ***Impact Factor: 5.156; ***2021 Impact Factor: 5.641;***
2020 Impact Factor: 5.355. *Indexing and Impact Factor URL: http://sjifactor.com/passport.php?id=21363*

DEVELOPMENT, STANDARDIZATION AND BENCH MARK OF MATHEMATICS PROFICIENCY TEST FOR SENIOR SECONDARY SCHOOL STUDENTS USING ITEM RESPONSE THEORY

Prof. Nkechi Patricia-MaryEsomonu[1]& Onyejiekwe Isaac Anayo[2]
[2]npm.esomonu@unizik.edu.ng[2], isaaconyejiekwe@yahoo.com
Department of Educational Foundations
Nnamdi Azikiwe University, Awka, Nigeria

ABSTRACT

In this study Mathematics Proficiency Test, (MPT) was for senior secondary school was developed validated and standardized. The researchers used an instrumentation design in the study. The researcher developed and administered one hundred (100) test item instrument titled Mathematical Proficiency Test (MPT) to a sample of 624 students drawn from six selected schools in the urban areas of Ahoada East Metropolis using proportionate sampling technique. A total of 600 copies of the instrument were returned to the researcher. After deleting the 5 respondents, that have missing items, the sample size was reduced to 595. The item parameters were calibrated using Mplus7. Based on the information, 9 items in the Mathematics Proficiency Test (MPT) were rejected as the items were too difficult. The researcher also found out that the discrimination parameters as presented, 36 items are within the range of 0.00 to 0.34 indicates very low discriminating power, while 54 items within the range of 0.35 to 1.34 indicate moderate discriminating power. Furthermore, 2 items had negative discriminating power. The study indicates that the instrument has reliability co-efficient of 0.7. The researcher recommended that mathematics teachers should use the instrument to measure mathematics proficiency level of students so that remedial help can be provided for them.

Keywords: Proficiency test, Mathematics, Item Response Theory, Standardization**.**

Introduction

Mathematics is one of the compulsory subjects in both Primary and secondary school education system in Nigeria. Mathematics is a language in which scientific ideas are expressed. It is a means by which other sciences including Physics, chemistry and Biology, and disciplines like Engineering and Geology are understood. Mathematics enables the various sciences to draw implications of their observation and experimental findings.

Despite the importance and role mathematics plays in the educational system, there has been some concerns on the performance of mathematics right from the inception of it in the Nigerian education. This is as a result of the fact that many students perform poorly in both internal and external examinations in mathematics (Wafaa, 2017). The researchers are of the opinion that this poor performance in mathematics may be connected with lack of students' proficiency in mathematics in which procedures and algorithms are taught in isolation without any recourse to training them in a conceptual understanding. In view of this, the National Council of Teachers of Mathematics (2001) has painstaking identified five processes by which children learn mathematics by doing

mathematics as follows: problem solving, reasoning and proof, representation, communication, and connections. Children that participate in the process of problem-solving build mathematical knowledge as well as understanding by dealing with and resolving genuine problems. They adopt reasoning and proof to create meaning out of mathematical tasks in addition to concepts and to cultivate, defend, and measure mathematical arguments and solutions. Moreover, children build as well as deploy representations such as diagrams, graphs, symbols, pictures, and manipulatives to think through mathematical challenges. They also involve in communication as they elucidate their ideas and thinking orally, in writing, as well as illustrations. Apart from evolving and deploying connections by children, they (children) also get new mathematical ideas and procedures but also make connections as well as linking mathematics and other disciplines by relating mathematics to real-world situations.

A lot of Nigerian students seem not to expand and show a profound understanding of and ability to do mathematics. This is against the Common Core State standards for Mathematics (2010) which highlighted the common eight Standards for Mathematical Practice which teachers should help children to develop in mathematics. They are making sense of problems and persevere in solving them; reason abstractly and quantitatively; construct viable arguments and critique the reasoning of others; model with mathematics; use appropriate tools strategically; attend to precision; look for and make use of structure; and look for and express regularity in repeated reasoning. According to Wu (2008) learning mathematics with meaningful understanding is a vital goal of mathematics teaching in schools. In addition, studies have shown that sharpening students' mathematical proficiency entails students learning mathematics with understanding (Carpenter and Lehrer, 1999; Hiebert, Carpenter, Fennema, Fuson, Wearne, Murray, Olivier, & Human, 1997; National research council (2001); Shafer & Romberg, 1999).

Mathematical proficiency can be defined as the quality of being skilled and showing expertise, competence, knowledge, beliefs, and facility in doing mathematics and becoming proficient problem solver with high productive disposition. On the other hand, NRC (2001 views mathematical proficiency as a composition of five interwoven and interdependent strands which includes conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. NRC (2001, p. 116) also defines conceptual understanding as the comprehension of mathematical concepts, operations, and procedures. A significant indicator of conceptual knowledge is being able to represent mathematical situations in different ways and knowing how different representations can be useful for different purposes" (NRC,2001, p. 119). It should be noted that representations allow students to see abstract mathematics concepts in diverse ways, which when intellectually structured and linked, support conceptual understanding. In conceptual understanding, students should be preoccupied with relational understanding-knowing what to do and why as opposed to instrumental understanding-knowing something by rote or without meaning and this relational understanding should be the

objective of daily teaching in mathematics. To achieve conceptual understanding, students ought to be made to witness the multiple entry points in solving a problem. Conceptual understanding allows students to build new knowledge as they make connections with earlier learned knowledge. According to Macgregor (2013), this method is far more useful to students than simple memorization of facts and procedures. NRC as cited by Awofala (2017) opines that conceptual understanding promotes retention and fosters the development of fluency.

According to (NRC, 2001, p. 121), Procedural fluency is defined as the knowledge of procedures, knowledge of when and how to use them appropriately, and skill in performing them flexibly, accurately, and efficiently". In the same vein, the National Council of Teachers of Mathematics (2000, p. 144) views procedural fluency as "having efficient, accurate, and generalizable algorithms for computing that are based on well-understood properties and number relationships. Hence, it is important to note that procedural fluency does not run contrary to conceptual understanding; in fact, the both work collectively to aid promote mathematical proficiency. In another development, Ostler (2011) states that, procedurally fluent students apparently develop the capability to appraise as well as simplify various expressions, work out simple equalities, as well as represent mathematical relationships in graphical form.

Strategic competence according to (NRC, 2001) is defined as the capacity to devise mathematical problems, represent them, and solve them. In a similar development, strategic competence is concerned with an individual's capacity to create a problem mathematically as well as employing his or her previous mathematical experiences to get a solution for it. Having strategic competence helps an individual to decode which strategies might be helpful in tackling the problem and in connecting these strategies to prior mathematical experiences. Students that lack enough strategic skill will often approach a mathematical problem with the purpose of using a trial-and-error strategy.

National Research Council (NRC) as cited in Awofola (2017) defined adaptive reasoning as "the capacity to think logically about the relationship among concepts and situations. Ability in adaptive reasoning helps an individual to consider alternative approaches, to follow the mathematical logic of a proposed pr oof, to note logical inconsistencies or contradictions, and to justify any conclusions (Sieffried, 2012). Students with adaptive reasoning are able to justify the solution steps employed in solving a problem in a logical manner in such a way that they know when the solution steps are wrong or right. Students are said to be capable of adaptive reasoning when they are able to think logically about the existing problems, estimating and reflecting through the problems and giving justifications for solving the problems.

On the other hand, Productive disposition can be defined as "the propensity to see sense in mathematics, to perceive it as both useful and worthwhile, to believe that steady effort in learning mathematics pays off, as well as to see oneself as an effective learner and doer of mathematics" (NRC, 2001, p. 131). Rather than see mathematics

as a set of arbitrary rules that an individual should memorise, students with fruitful disposition view mathematics as a system of connected conceptions that can be understood with perseverance and diligent effort. NRC further stated that a line of distinction can be drawn between this strand and the other four strands although productive disposition is required to build the other four strands while the other four strands deal with an individual's cognitive processes and relate to mathematical content knowledge, productive disposition is enshrined in an individual's affect, beliefs, and identity and strengthening of the other four strands assists build an individual's productive disposition. Thus, a symbiotic relationship exists between productive disposition and the other four strands. Students who do not possess productive dispositions might not see themselves as learners or doers of mathematics and might not see value in steady effort and perseverance in mathematics. In addition, these students might not even think that mathematics is supposed to make sense.

It can be said that developing mathematical proficiency involves getting the ability to engage in mathematical habits of mind that encourage not only procedural fluency other than conceptual understanding, adaptive reasoning, and strategic competence inside the confine of mathematisation. While mathematisation plays a significant role in developing proficiency, the processes of mathemitisation which include connecting, communicating, reasoning arguementation , justifying , representing , problem- solving and generalizing ought to infiltrate all learning and teaching activities in mathematics (NCCA , 2014)

In Nigeria, mathematics is viewed as a problem-solving activity and this view is ingrained in the new senior secondary school mathematics curriculum. But evidence suggests that there is a disparity among this view and mathematics that the majority of students are exposed to in the classroom. Hence, it is obvious that most teachers adopt the traditional lecture method in teaching mathematics (Ifamuyiwa & Akinsola, 2008); Awofala, Fatade & Ola-Oluwa, 2012) and more often than not teachers are glued to the traditional mathematics textbooks when giving instruction in mathematics with emphasis on lower-order thinking skills (Awofala, 2012) and mathematical procedures. These teachers' activities negate not only the principle of constructivism but deny students the chance to experience mathematics as a creative and engaging process. In general, researches into mathematical proficiency are in its infancy stage despite its importance as a goal that must be nurtured in the mathematics classroom. In another development, Wu (2008) assessed 491 Chinese sixth graders" mathematics proficiency as reflected in conceptual understanding, procedural fluency, and competence in word problem applications using the model – strategy - application (MSA) approach. The results revealed that Chinese students" procedural fluency was at a higher level compared to their conceptual understanding and word problem in real-world applications. Furthermore, the result revealed that a higher level of computation did not lead Chinese students to a deep understanding of fractions and decimals. Similarly, Kinnari (2010) investigated the mathematical proficiency of first year engineering students at Tampere University of Applied

Sciences (TAMK) in which TAMK had designed a diagnostic procedure based on the NRC's five strands of mathematical proficiency framework. The students took a 20 multiple question online survey based on three of the strands: conceptual understanding; procedural fluency; and strategic competence. The results revealed that the students scored from 8.1 to 18.3 points (out of 40), with a mean score of 13.8, representing a significant lack of mathematical proficiency. On the other hand, Samuelson (2010) investigated the effect of two differently structured methods, traditional and problem- solving on students' mathematical proficiency in Sweden. The results also showed that there were no significant differences between teaching methods when assessing procedural fluency but students' progress in conceptual understanding, strategic competence, and adaptive reason was significantly better when teachers taught with problem -based curriculum. In addition, students' productive disposition, beliefs in diligence and one's own efficacy, was affected significantly more if students worked traditionally.

According to Samuelson (2010), taking mathematical proficiency as the purpose of mathematics education has the probability to transform the kind of mathematics and mathematical learning that young children are exposed to (NCCA,2014) irrespective of their gender. Unlike the industrialized nations of the world, where researches into mathematical proficiency had reached an appreciable level, there were scarceness of studies in Nigeria on students' mathematical proficiency and mathematical proficiency related issues.

A good test should be reliable and valid. A combination of these is the entire validation process. This will of necessity include test validity, test reliability as well as general item analysis. Furthermore, a good item analysis will include determination of the item difficulty, discrimination as well as the item distracter. According to Kpolovie (2016: 337), reliability is the extent which a measuring instrument accurately measures a particular social or psychological trait, construct, characteristics or behaviour of the examinees in the best perfect manner without any unsystematic (random) or systematic fluctuation (error) both in the instrument and in the characteristics under investigation. Reliability checks involves many approaches. Huck (2007) stated that developing a valid and reliable instrument usually requires multiple iterations of piloting and testing which can be resource intensive.

A clear majority of standardized proficiency test were developed based on Classical Test Theory (CTT), which is a useful theory for understanding latent traits (Diener et al., 2018) for measures. Brown et al., (2018) and Oishi (2007) in their separate studies reported that evaluations of psychometric information of items and scales using CTT is dependent on the number of items and on the sample's size and other features, so any changes of these features can strongly affect both item and the total psychometric properties of the scale. For instance, more precise estimates of reliability coefficients and their confidence intervals are obtained in large sample sizes of at least 400 respondents (Charter, 1999), which is not so common when these measures have been tested (Leue and Lange, 2011).

IRT is as relatively modern psychometric technique that overcomes some of these limitations. One of IRT's biggest advantages is that we can determine how suitable items are to measure the latent traits, so it can increase reliable information and validity of the scale as a whole. The error and the reliable information obtained using IRT vary from one item to another and throughout the trait continuum of the scale, sometimes widely for one part of the scale compared with other parts (Oishi, 2007).Indeed, IRT models provide item parameters that enable the evaluation of how well an item performs in measuring the underlying construct. More specifically, IRT is a model that provides a linkage between item responses and the latent characteristic assessed by a scale. IRT assumes that each examinee responding to a test item possesses some amount of the underlying ability (denoted by the Greek letter theta).

Assumptions of item response theory include dimensionality which is described as the number of traits underlying a test that accounts for variation in students' test performance. The student's performance, in an extremely set of n-dimensional latent space are often described by a vector of ability scores as (θ1, θ2, θ3…. θn). Item response models that presume many latent traits which accounts for students' test performance is known as multidimensional item response model, whereas item response models that assume a single latent ability is known as unidimensional models (Comrey & Lee,1992; Hambleton& Swaminathan,1985). This implies that the test items assess single area of construct or knowledge or ability. A unidimensional test comprises of items that tap into only one dimension (Adewale, 2017; Bejar,1980& Lynch,1985). Anywale, Isaac-Oloniyo and Abayomi (2020) study on assessment of dimensionality of test item revealed that Osun State unified Mathematics Achievement Test items is multidimensional in nature. Findings from the work of Oguoma, Metibemu and Okoye (2016) on dimensionality assumption test on 2014 Mathematics achievement items of West African Senior Secondary Certificate Examination (WASSCE) concluded that the test items of WASSCE mathematics were inherently multidimensional in nature. Furthermore, Okwilagwe and Ogunrinde (2017) also found that fifty (50) items of 2013 WASSCE and sixty (60) items of National Examinations Council (NECO) Geography respectively violated assumption of unidimensionality and that there was more than one dimension that accounted for the variation observed in examinees to the geography test items.

Item local independence is one of the concepts of Item Response Theory (IRT), which has received great attention by researchers and authorities in the area of psychometrics because of its importance in probability theory. Yen (1993) identified two forms of independence, namely: statistical independence and stochastic independence. Statistical independence which is important for this present study refers to a situation in which two quantum systems acting randomly are said to be prepared differently. That is, when items are statistically independent, each exhibits its quality and it takes examinees' good display of ability in unfolding the characteristic function about them. Based on this assertion, Ayang (2004) posited that, in order to set mathematics test items that would not violate local independence, the interaction between each test constructs

(items) must not be high, but as low as possible with respect to logic operations and manipulations. Items that tend to have the same pattern but with different ability levels should be distributed randomly to span the entire tests length. He concluded by recommending the avoidance of chain items since they can give clue to one another. Several studies carried out using different standardized achievement tests like the University Matriculation Examination (UME), the West African Senior Secondary Certificate Examination (WASSCE) and the National Examination Council Senior Certificate Examination (NECO/SSCE) reveal that the stability of performance across test takers can be accountable for by the level of independence of items in the test. Ubi (2006) assessed 800 scripts of candidates who wrote Mathematics in the University Matriculation Examination in Cross River State, Nigeria and discovered that a good number of items in the examination were locally independent. Thissen (1991) and Sireci, Thissen and Wainer (1991) have separately pointed out that standardized examinations are basically known to have high local items independence because of some basic principles of testing formally adhered to in conducting the examinations. They pointed out that an appropriate assumption of local item dependence will produce over estimates of test information and reliability and underestimates of the standard error of estimate. Thissen and his colleagues have examined the possible effect of local item independence Although a number of different IRT models exist, the most commonly employed one is the two-parameter logistic model (2PL), which assumes a single underlying ability and two item parameters: a difficulty parameter (b) and a discrimination parameter (a).

The discrimination parameter is aka the α parameter. It is used to determine how well the items discriminate against different levels of the latent trait. In this case, ranging from -4 to 4. The αiαi parameter another name is the slope parameter. Steeper slopes at particular level of the traits indicate that it is more discriminative than levels of the traits with gentler slopes. The most discriminative point is always at the centre of the curve. In theory, the values for the α ranges from −∞−∞ to +∞+∞. Negative values of the αiαi are possible but considered problematic as it would mean that items with increasing levels of ability are less likely to endure more severe response options. This could happen when the item poorly discriminates between levels of the ability (illogical relationship) or there was some sort of coding error. The discrimination values (parameter a-values) of good items according to Adedoyin and Mokobi (2013) ranges from 0.5 to 2 . If the values of the item discrimination a is above 1, this is normally desirable value for a good test item. Baker (2017) describes the range and interpretation of values for item discrimination as follows: very low; .01 - .34, Low; .35 - .64, moderate; .65 - 1.34 High; 1.35 - 1.69 and Very high; 1.70 and above.

The difficulty parameter is used to describe how difficulty an item is to achieve a 0.5 probability of correct response at a given ability. Therefore, if the 50% chance of correctly answering an item is at the right side of the plot, then it means that a higher level of ability is needed to achieve that goal. Baker (2017) pointed out that theoretically, difficulty values can range from - 00 to + 00, in practice, difficulty values usually are in the range of - 3 to + 3.

Ceniza and Cereno (2012) gave the interpretation for values of item threshold/difficulty (b-values) as follows: Very Easy = Less than -2, Easy = -0.50 to -2.00, Average = -0.49 to 0.49, Difficult = 0.50 to 2.00 and Very Difficult = Greater than 2.00

The guessing parameter is more widely used in the educational setting where guessing the correct answer for an item may be attempted by the test taker. So, the IRT model has to account for the guessing or the pseudo-chance-level parameter. An examinee who randomly selects responses to items of four options can answer these items correctly about 1 out of 4 times, meaning that the probability of guessing correctly is about 0.25 (Eleje & Esomonu, 2018). Harris (1989) concluded that the items with 0.30 or greater c-values are considered not very good, rather c-values of 0.20 or lower are desirable.

Measures of model fit and parameter estimates are obtained through maximum likelihood estimation. In short, there is paucity of mathematics proficiency test in secondary schools. Hence, the aim of the present study is to apply IRT in development and validation of mathematics proficiency test for secondary school students.

The following research questions guided the study:

1. To what extent are the assumptions of item response theory met in calibration of item parameters of the mathematics proficiency test?
2. Which model of item response theory is most appropriate for the calibration of item parameters of the mathematics proficiency test?
3. How valid are the items to the model of item response theory appropriate for the calibration of item parameters of the mathematics proficiency test?
4. What are the item parameter estimates of the mathematics proficiency test?
5. How reliable is the mathematics proficiency test for secondary school students?

Method

The researcher used instrumentation design in the study. The researchers developed and administered one hundred (100) test item instrument titled Mathematical Proficiency Test (MPT) to a sample of 624 students drawn from six selected schools in the Urban areas of Ahoada East Metropolis. Before the actual administration, the researcher did a preliminary validation by given it to three mathematics experts to check its validity and correctioness of the keys and as well as undertook a pilot test using 60 respondents to determine the reliability of the MPT. The experts scrutinized the test questions and their keys and finally certified the instrument as valid. The 624 actual samples for the study were drawn using proportionate sampling technique. The researchers carefully with the help of two research assistants distributed the instruments to the respondents in the six selected schools in the Ahoada East Metropolis. Administration of the instrument was done over a period of three (3) weeks. After the responses, copies of the instrument were collected on the spot and a total of 600 out of 624 copies were successfully completed, as 24 copies were not returned due to non-response/ incomplete response by the respondents. Hence, the total sample size

used for the analysis was 600. Nkech Esomonu proficiency bench mark of 90% and above was used in interpreting the students' proficiency level. In term of dimensionality, DIMTEST statistics in DIMPACK software was used in this study.

Preliminary data analysis was carried out in terms of data screening for missing values, kurtosis and skwenness. In item of missing values, 5 respondents had missing data representing 0.83% , hence litewise deletion approach was adopted. After deleting the 5 respondents, the sample size was reduced to 595. Thereafter, the kurtosis and skewness of the data were examined to ascertain if the data were normally distributed. Results of kurtosis and skewness indicate that 8 items (Items 4,12,16,57,61,66,67 and 99) exhibit extreme kurtosis and skewness since their values were above -7 and +7 as recommended (Tabachnick & Fidell, 2012). The aforementioned items were deleted and this reduced the number of items on the test to 92. Normal Q-Q plot was then conducted to ascertain if they satisfied the assumption of normality. Figure 1 show that normality was not grossly violated after the 8 items were removed, hence the choice of maximum likelihood estimation of parameters.

DIMTEST statistic of essential unidimensionality. The null hypothesis posits that the assessment subset (AT) and partitioning subset (PT) partitions assess the same dominant underlying dimension, while the alternative hypothesis implies that the items in the AT partition are best represented by a dimension that is distinct from that driving responses to the PT items. To examine the number of dimensions, parallel factor analysis was conducted. Parallel Analysis is arguably the preferred technique right now for determining number of factors to extract (Anywale, Isaac-Oloniyo & Abayomi, 2020). Parallel Analysis essentially simulates data that is similar to yours but where there is less commonality between items. It then uses this simulated data to provide reasonable suggested cut-offs for your scree plot (that you would normally make just eyeballing the scree plot). It provides a more objective evaluation of the scree plot than the eye test.

Parallel analysis figure 2 suggests that the number of factors = 22 and the number of components = 16. The implication is that the mathematics proficiency test is multidimensional. In other words, different aspect of mathematics proficiency was measured local independence assumption of item response theory was investigated based on Yen's Q3, residuals for any pair of items should be uncorrelated, and generally close to zero. Residual correlations that are high indicate a violation of the local independence assumption, and this suggests that the pair of items have something more in common than the rest of the item set have in common with each other (Marais, 2013). For this study, using Yen's Q3 to screen items for local dependence, no item residual correlations below absolute value of 0.2 for the Q3 statistic itself was found. This indicates that the assumption local independence of item response theory was not violated.

To answer research question two, the examinee responses were also subjected to full information item factor analysis compared using (Akaike Information Criterion (AIC), and Likelihood ratio test (LogLik) in order to establish the best fit model that provided the information for the

exact number of factors embedded in the test data. The fitness of hypothesised models was assessed using the information criteria with the most parsimonious model. To select appropriate item response theory model for calibrations, 1-parameter, 2-parameters and 3-parameters model of item response theory were run using Mplus 7. The researcher tested the item fit statistic using the Orlando–Thissen–Bjorner *item fit S-χ2 statistic* to determine absolute fit of the model to each item. Regarding S-χ2 statistic, a value that is not significant indicates that the model adequately fits an item.

Huck (2007) considered a reliability co-efficient of 0.7 and above for instrument to be considered reliable for a particular study especially for standardized instrument. Only the items with poor difficulty values and poor discriminating values will be rejected and those with negative discriminating vales. Ceniza and Cereno (2012) gave the interpretation for values of item threshold/difficulty (b-values) as follows: Very Easy = Less than -2, Easy = -0.50 to -2.00, Average = -0.49 to 0.49, Difficult = 0.50 to 2.00 and Very Difficult = Greater than 2.00.

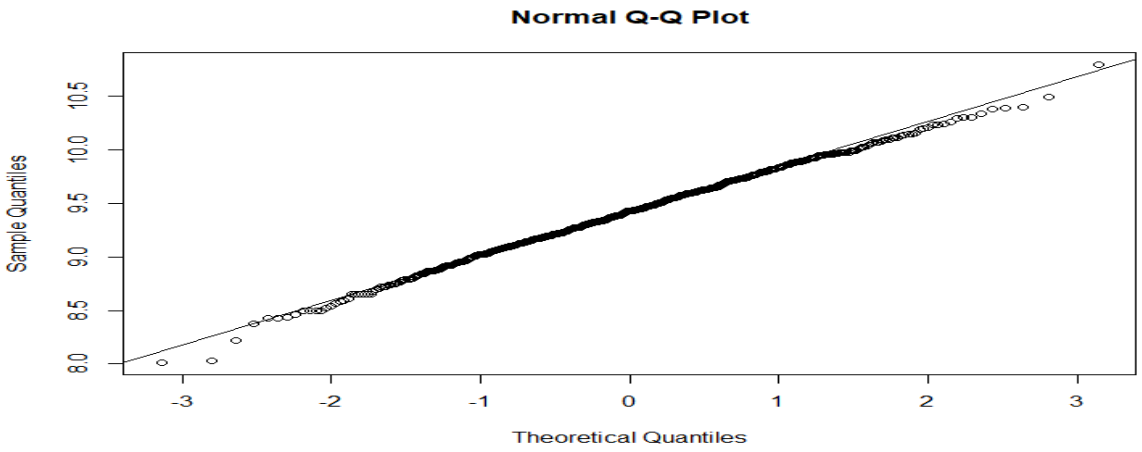
Results

Preliminary Data Analysis

Preliminary data analysis indicated that only 5 respondents had missing data representing 0.83%. After deleting the 5 respondents, the sample was reduced to 595. Thereafter, the kurtosis and skewness of the data were examined to ascertain if the data were normally distributed. Results of the kurtosis and skewness indicate that 8 items (i.e., 4, 12 16, 57, 61, 66, 67 and 99) exhibit extreme kurtosis and skewness. The aforementioned items were deleted and this reduced the number of items on the test to 92. Normal Q-Q plot was then conducted to ascertain if they satisfied the assumption of normality.

Figure 1 show that normality was not grossly violated after the 8 items were removed, hence the choice of maximum likelihood estimation of parameters.

Figure 1: Normal Q-Q plot of Mathematics Proficiency Test



Research Question 1: To what extent are the assumptions of item response theory met in calibration of item parameters of the Mathematics Proficiency Test?

UNIJERPS
Unizik Journal of Educational Research and Policy Studies VOL.19 (3);
January-June, 2025 https://unijerps.org
2023 Impact Factor:* 6.834;** 2022 ***Impact Factor: 5.156; ***2021 Impact Factor: 5.641;***
2020 Impact Factor: 5.355. *Indexing and Impact Factor URL: http://sjifactor.com/passport.php?id=21363*

Assumptions of item response theory- dimensionality

Table 1

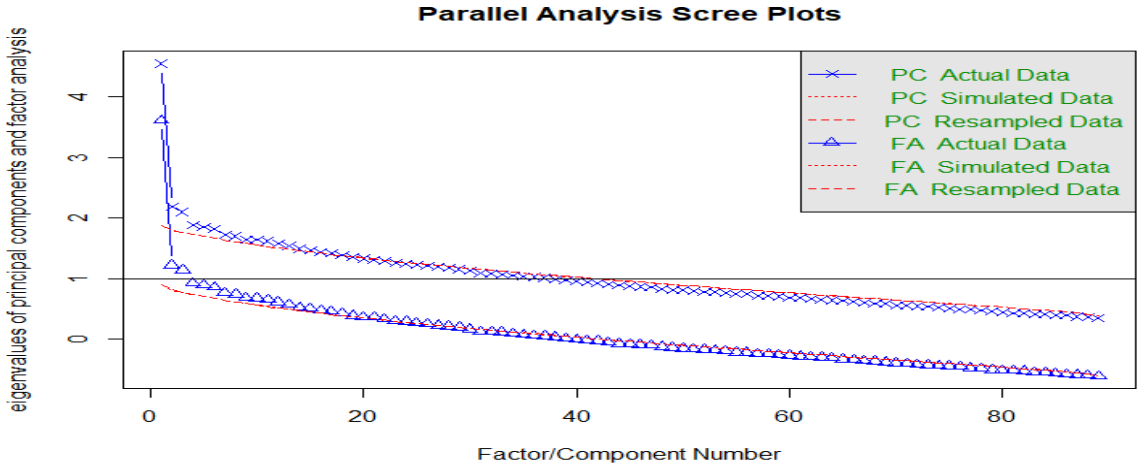
Stout's Test of Essential Unidimensionality Test

TL	TGbar	T	p-value
8.4823	5.2447	3.2215	0.0006

Table 1 revealed that Mathematics proficiency test items violated essential unidimensionality assumption with (T= 3.2215, p<.05). This led to the conclusion that the AT items were dimensionally distinct from the remaining items in PT. Therefore, multidimensionality was manifest.

Parallel analysis to determine the number of factors.

Figure 2: Parallel analysis of Mathematics Proficiency Test



Parallel analysis figure 2 suggests that the number of factors = 22 and the number of components = 16. The implication is that the mathematics proficiency test is multidimensional. In other words, different aspect of mathematics proficiency was measured by the instrument.

Test of Local Independence assumption showed that no item residual correlations below absolute value of 0.2 for the Q3 statistic itself was found. This indicates that the assumption of local independence of item response theory was not violated.

Research Question 2: Which model of item response theory is most appropriate for the calibration of item parameters of the Mathematics Proficiency Test?

The results of Model fit of various model are presented in table 2 for comparison.

Table 2: Model Fit Statistics of Instrument

Model	Loglik	AIC
1PL	-37099.287	74384.574
2PL	-36928.128	74224.257
3PL	-37013.294	74578.589
4PL	-39510.681	76615.360

From Table 2, the 2-parameter model of item response theory was most appropriate for calibrating item parameters. The number of items fit per model is also presented in Table 3. The result corroborates the findings that the two parameters model is most appropriate for the calibration of the mathematics proficiency test.

Research Question 3: How valid are the items to the model of item response theory appropriate for the calibration of item parameters of the Mathematics Proficiency Test?

Table 3: Item fit of the Mathematics Proficiency Test to Various Models

Model	No. of item fit	No.of misfit item	Percentage of item fit
1PL	80	12	86
2PL	85	7	92
3PL	83	9	90
4PL	78	14	84

The results in table 3 indicates that 85 items representing 92% of the Mathematics Proficiency Test items fitted the 2-parameter model since the Chi-square p-value is greater than 0.05. Therefore, majority of the items of the test are valid for calibration using the 2-parameter logistic model of item response theory than the other models of item response theory for this study.

Research question 4: What are the items parameters estimates of the mathematics proficiency test?

The results for difficulty and discrimination item difficulties are presented in tables 4 and 5 respectively.

Table 4: Item difficulty (b) parameter estimate of the Mathematics Proficiency Test (MPT)

Domains	Item	B	Domains	Item	B
Logarithm	46	-1.429		20	-0.901
	47	-0.322		42	**-11.022**
	48	-0.396	Algebraic fraction	3	-0.449
	49	0.105		68	-0.021
Approximation	1	0.027		84	-0.317
	2	0.773		85	-0.261
	15	-0.462		91	-1.440
	30	-0.989		92	0.261
	45	0.060		93	-0.793
	78	-0.105		94	0.511
	86	-0.569		95	-0.381
	87	-0.142		96	-0.465
	88	-0.011		97	**6.368**
	44	-0.175		98	-0.642
Sequence & series	58	-0.045	Circle geometry	65	**-9.295**
	59	-0.304		40	-0.357
	60	-0.518		41	-0.315
	74	0.093		51	-0.296
	73	-0.502		52	0.080
	75	-0.134	Trigonometry	35	-1.549
	76	**-4.191**		36	-0.556

	77	-0.241		37	-0.529
Quadratic equation	22	-0.158		38	-0.138
	23	-0.201		39	-0.332
	24	-0.338	Bearing	43	-0.168
Quadratic equation	25	-0.548		53	-0.111
	26	**8.742**	Measures of Tendency	11	-1.668
	27	-0.313		13	-0.262
	28	-0.550		14	0.148
	29	-1.073		70	-1.708
	62	-0.197		31	-1.012
	81	-0.226		32	0.037
	82	0.068		33	-0.192
Simultaneous equation	21	-0.214	Measures of dispersion	8	-0.237
	55	0.060		9	-0.648
	56	-0.252		10	-0.384
Gradient curve	63	-0.041		83	-0.401
	64	-1.930		50	0.148
	79	-0.165		100	**-3.702**
	80	0.295	Statistical graph	34	-0.894
	69	-0.105		72	-0.084
Logic reasoning	89	-0.461	Probability	5	**5.013**
	90	-0.519		6	**-3.702**
Linear inequalities	17	-0.585		7	-0.179
	18	-1.131		54	-0.378
	19	-0.639		71	-0.770

Table 4 indicates that 74 items fall between -3 to 0 while 10 items fall between 0 to +3 on the logit scale. 8 items fall outside -3 to + 3 logit scale. Based on the information, the 8 items in the Mathematics Proficiency Test (MPT) were rejected as the items were too difficult. The 8 items are boldly indicated on Table 4.

Table 5: Item discrimination (a) parameter estimate of Mathematics Proficiency Test

Domains	Item	A	Domains	Item	A
Logarithm	46	**0.203**		20	**0.311**
	47	0.467		42	**0.040**
	48	0.357	Algebraic fraction	3	0.361
	49	0.377		68	0.662
Approximation	1	**0.027**		84	0.450
	2	0.773		85	**0.168**
	15	0.465		91	**0.192**
	30	0.356		92	0.550
	45	0.436		93	**0.149**
	78	0.354		94	0.462
	86	0.535		95	**0.272**
	87	**0.310**		96	0.380
	88	0.614		97	**-0.048**
	44	**0.311**		98	**0.261**
Sequence & series	58	0.777	Circle geometry	65	**0.045**
	59	0.397		40	0.617
	60	0.531		41	0.527
	74	0.345		51	**0.311**
	73	**0.155**		52	0.411
	75	**0.278**	Trigonometry	35	**0.242**
	76	**0.062**		36	0.463
	77	**0.268**		37	0.444
Quadratic equation	22	0.504		38	0.867
	23	0.656		39	0.497
	24	0.377	Bearing	43	0.431
Quadratic equation	25	0.414		53	0.523
	26	**-0.055**	Measures of Tendency	11	**0.216**
	27	0.556		13	**0.272**
	28	0.427		14	0.529
	29	**0.306**		70	**0.232**

	62	0.715		31	**0.212**
	81	**0.286**		32	0.599
	82	0.511		33	0.379
Simultaneous equation	21	0.471	Measures of dispersion	8	0.432
	55	0.633		9	0.404
	56	0.577		100	**0.101**
Gradient curve	63	0.542		14	0.529
	64	**0.212**		50	0.468
	79	**0.225**		10	0.407
	80	**0.318**	Statistical graph	34	0.379
	69	**0.288**		72	0.691
Logic reasoning	89	0.448	Probability	5	**-0.146**
	90	**0.324**		6	**0.175**
Linear inequalities	17	0.560		7	0.484
	18	**0.195**		54	0.578
	19	**0.307**		71	0.309

The discrimination parameters are presented in Table 5. 36 items are within the range of 0.00 to 0.34 which indicates very low discriminating power, while 54 items are within the range of 0.35 to 1.34 which indicate moderate discriminating power. Furthermore, 2 items had negative discriminating power. Based on the report on discrimination parameters estimate on Table 5, 8 items (ie items 42, 97, 65, 76, 26, 100, 5 and 6 boldly typed were rejected.

Research question 5: How reliable is the mathematics proficiency test for secondary school students?

Figure 3: Latent Distribution Mathematics Proficiency Test

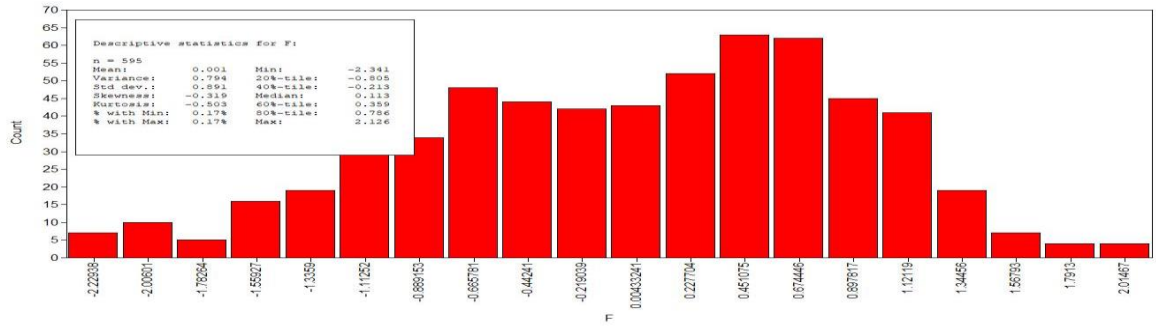
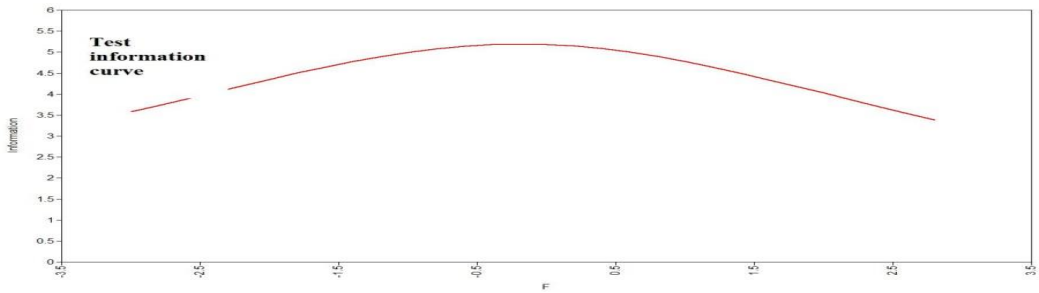


Figure 4: Test information Curve of Mathematics Proficiency Test



In terms of reliability, it was conducted using item response theory approach, Figure 3 and 4 indicated that vary levels of abilities in mathematics were covered by the instrument. The empirical reliability of

the instrument was found to 0.79. This indicates that the instrument is reliable.

Discussion

It was found out that the mathematics proficiency test was multidimensional. In other words, different aspects of mathematics proficiency were measured by the instrument. The finding agreed with that of Anywale, Isaac-Oloniyo and Abayomi (2020) on assessment of dimensionality of Osun State Unified Mathematics Achievement Test items, who found that the test was multidimensional in nature. It also agreed with the findings of Oguoma, Metibemu and Okoye (2016), who in their study on dimensionality 2014 Mathematics achievement items of West African Senior Secondary Certificate Examination (WASSCE), concluded that the test items were inherently multidimensional in nature. Furthermore, Okwilagwe and Ogunrinde (2017) also found that each of the 50 - item 2013 WASSCE and 60-item of National Examinations Council (NECO) questions in Geography violated assumption of unidimensionality, implying that there was more than one dimension that accounted for the variation observed in examinees to the geography test items.

In terms of local independence, it was also found out that the assumption was ascertained using Yen Q3 statistic was not violated. The finding also agreed with that of Ubi (2006) who assessed 800 scripts of candidates who wrote Mathematics in the University Matriculation Examination in Cross River State, Nigeria and discovered that a good number of items in the examination were locally independent. Thissen (1991) and Sireci, Thissen and Wainer (1991) have separately pointed out that standardized examinations are basically known to have high local items independence because of some basic principles of testing formally adhered to in conducting the examinations. They pointed out that an appropriate assumption of local item dependence will produce over estimates of test information and reliability and underestimate of the standard error of estimate.

It was also found out that 73 items fell between -3 and 0 while 10 items fell between 0 to +3 on the logit scale. 8 items fell outside -3 to + 3 logit scale. Based on the information, 8 items in the Mathematics Proficiency Test (MPT) were rejected as the items were too difficult. This is in line with Baker (2017) who pointed out that theoretically; difficulty values can range from - 00 to + 00, in practice, difficulty values usually are in the range of - 3 to + 3.

The researchers also found out that the discrimination parameters as presented, 36 items were within the range of 0.00 to 0.34 which indicated very low discriminating power, while 54 items fell within the range of 0.35 to 1.34 which indicated moderate discriminating power. Furthermore, 2 items had negative discriminating power. Based on the report on discrimination parameters estimate, 8 were rejected. The above findings are in line with the findings of Adedoyin and Mokobi (2013) whose estimated ranged from 0.5 to 2. If the values of the item discrimination are above 1, this is normally desirable value for a good test item. When the reliability of the test was ascertained using item response theory approach, empirical reliability of the instrument was found to be 0.76. This indicates that the instrument was reliable. Huck (2007) considered a reliability co-efficient of 0.7 and above for instrument to be considered reliable for a particular study.

Conclusion

The final instrument was made of 54 items, based on the results of the analysis. The 54 items are valid and reliable, and could be used to assess Mathematics Proficiency Test of secondary schools in Ahoada East Local Government of River State

Recommendations

Based on the findings, the researcher recommends the following:

1. Mathematics teachers should use the instrument to measure mathematics proficiency level of students so that remedial help can be provided for them.
2. The secondary school counselors should use the Mathematics proficiency instrument developed based on item response theory for assessment of students' proficiency in learning mathematics during selection and placement of students in school.
3. The instrument should be administered to the students during classroom instructions and feedback information provided to help the students to set specific learning goals and improve learner autonomy in order to master relevant skills that will help them perform well in mathematics.
4. The Nigerian Mathematical Centre should adopt the instrument to measure mathematics proficiency level of students.

REFERENCES

Adedoyin, O.O. & T. Mokobi, (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4),992-1011.

Adewale, M.A. (2017). A training manual on item response theory, (1st ed.) Ibadan: Institute of Education, University of Ibadan in Collaboration with National Examinations Council, Minna, Niger State. *Int J Eval& Res Educ*., 2252-8822

Awofala, A.O.A, (2012). An Analysis of the new –year basic Education mathematics Curriculum in Nigeria. Acta Didactica Napocensia, 5(1), 17-27

Awofala, A.O.A, Fatade, A.O. & Ola-Oluwa, S.A. (2012). Achievement in Cooperative and individualistic goal – structured Junior Secondary School mathematics, classrooms in Nigeria. *International journal of mathematics Trends and Technology, 3(1), 7-12*

Awofala, A.O.A. (2017). Assessing senior secondary students' mathematical proficiency as related to gender and performance in mathematics in Nigeria. International Journal of Education and Science (IJRES), 3(2), 488-502. DOI: 10.21890/ijres.327908

Ayang, E. E. (2004). *Item local independence and students' mathematics ability in Calabar.* Unpublished M.Ed Thesis, University of Calabar, Nigeria.

Baker, F.B. (2017). *The basics of item response theory.2nd Edn.* United States of America: ERIC Clearinghouse on Assessment and Evaluation.

Bejar,I.,(1980).A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *J. Educ. Meas*., 17(4),283–296.

Brown, N. J. L., Lomas, T., & Eiroa-Orosa, J. (2018). *The Routledge international handbook of critical positive psychology*. New York: Routledge.

Carpenter, T.A & Lehrer, R. (1999). Teaching and Learning Mathematics with Understanding. Editors: Fennema, E. And Romberg, T. A.

Ceniza, J.C. & Cereno,D.C.,(2012). Development of mathematics diagnostic test for DORSHS.Available.http://www.doscst.edu.ph/index.ph[/academics/graduateschool/publication/category/5-volum-1-issue-1-2012

Charter R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *J. Clin. Exp. Neuropsychol*, 21, 559–566.

Comrey, A. L. & Lee, H. B. (1992).*A first course in factor analysis.* Hillsdale, NJ: Erlbaum.

Common Core State Standards initiative (CCSSI). (2010). Cross – national patterns of gender differences in Mathematics: A Meta – analysis, Psychological Bulletin, 36(1), 103 – 127.

Diener, E., Lucas R. E., & Oishi S. (2018). Advances and open questions in the science of subjective well-being. *Collabra Psychol.* 4,15.

Eleje, L. I & Esomonu, N. P.M. (2018). Test of achievement in quantitative economics for secondary schools: Construction and validation using item response theory. *Asian Journal of Education and Training*, 4(1),18-28.

Hambleton, H.& Swaminathan, R.K. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hiebert, J. Carpenter, T., Fennema, E., Fuson, K., Wearne, D., Murray, H., Olivier, A., & human, P. (1997). Making sense: Teaching and learning mathematics with understanding, Portsmouth, NH: Heinemann.

Ifamuyiwa, S.A & Akinsola, M.K. (2008). Improving Senior Secondary school, students' attitude towards mathematics. *International Journal of mathematical Education in Science and Technology, 39(5), 569 – 585*

kinnari, H. (2010). A study of mathematics proficiency. Scientific Bulletin. Serires, Mathematical Modelling in Civil Engineering, 23-27. retrieved from Research Library.

Kpolovie, P.J. (2016). Excellent research methods. Indians, USA: Patridge Publishing

Leue A., & Lange S. (2011). Reliability generalization: An examination of the positive affect and negative affect schedule. *Assessment*, 18, 487–501.

Lynch, B. (1985). *The factor structure of the Winter*1985 ESLPE. University of California at Los Angeles, 1985.

MacGregor, D. (2013). Developing mathematical proficiency. EPS, Literacy and Intervention, Academy of Math.

Marais, I. (2013). Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 111-130). London, England: Wiley

National Council for Curriculum and Assessment (2014). Mathematics in early childhood and Primary education (3-8 years). Research report No. 18, 35 Fitzwilliam Square, Dublin ,2.

National Council of Teachers of Mathematics (2000). Principles and standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

National Research Council (2001). Adding it up. Helping Children learn mathematics. J. Swafford, and B. Findell (Eds). Mathematics Learning Study Committee, Center for Education, Division of Behavioural and Social Sciences and Education. Washington, DC: National academy Press. Retrieved from http:://www.nap.edu/catalog/9822. html.

Oguoma, R. O., Metibemu, C.C. &Okoye, M.A. (2016). An assessment of the dimensionality of 2014 West African secondary school examination mathematics objective test scores in Imo State, Nigeria, *African J. Theory Pract. Educ. Assessment*, 4, 18-33.

Oishi S. (2007). The application of structural equation modeling and item response theory to cross-cultural positive psychology research, in *Series in Positive Psychology. Oxford In Ong* A. D., van Dulmen M. H. (Eds) M. (New York, NY: Oxford University Press;), 126–138. *Handbook of Methods In Positive Psychology*.

Okwilagwe, M. A.& Ogunrinde, E.A. (2017). Assessment of unidimensionality and local independence of WAEC and NECO 2013 geography achievement tests. *African J. Theory Pract. Educ. Assess.*,5,31-44.

Osegbue, G. C. (2025). Early childhood education in Nigeria: challenges and prospects. *Unizik Journal of Educational Laws and Leadership Studies*, *2*(1).

Osegbue, G. C. (2025). Crisis intervention and response in educational institutions: developing and implementing plans in developing countries. *African Journal of Educational Management, Teaching and Entrepreneurship Studies*, *15*(1).

Ostler, E. (2011). Teaching adaptative and strategic reasoning through formula derivation: beyond formal Semiotics, Sutra. *International Journal of Mathematics Science Education, 4(2), 16-26.*

Samuelsson, J. (2010). The impact of teaching approaches on students' mathematical proficiency in Sweden. *International Electronic Journal of Mathematics Education, 5(2), 61-78.*

Shafer, M. & Romberg, T. (1999). Assessment in Classroom that promote understanding (pp. 159 – 184). Mahwah, NJ. Eribaum

Siegfried, J.M. (2012). The hidden Strand of mathematical proficiency: defining and assessing for productive disposition in elementary school teachers in mathematical content knowledge.

An unpublished Ph.D. Thesis, University of California, San Diego and San Diego state University, USA.

Sireci, S. G., Thissen, D. &Wainer, H. (1991). On the reliability of test lets based tests. *Journal of Educational Measurement*, 28, 237-240.

Tabachnick,B. G., & Fidell, L. S. (2012).Using Multivariate Statistics. 6th Edition, Pearson Education, Boston.

Thissen, D. (1991). *Multilog users' guide: Multiple categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software International.

Ubi, I. O. (2006). *Item local independence, dimensionality and trend candidates' mathematics performance on University Matriculation Examination in Nigeria.* Unpublished Ph.D Dissertation, University of Calabar, Nigeria.

Wu, Z. (2008). Using the MSA Model to assess Chinese Sixth Graders' Mathematics Proficiency. Journal of Mathematics Education. Vol.1, No.1, pp. 74-95

Wu, Z. & An, S. (2008, July). Using the Model – Strategy- application approach to developing Pre-service teachers '. Knowledge and assessing thesis progress in Math. Methods courses. Paper presented at 11th International Congress on Mathematics Education, Monterrey, Mexico.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement,* 30, 197-213.