WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Development of Mathematics Achievement Test Using Item Response Theory.

Isaac Anayo Onyejiekwe[1], Romy O. Okoye[2]

[1]Department Of Educational Foundations, Faculty Of Education , Nnamdi Azikiwe University , Awka Isaaconyejiekwe@yahoo.com
[2]Department Of Educational Foundations, Faculty Of Education , Nnamdi Azikiwe University , Awka .
DOI : https://doi.org/10.55248/gengpi.5.1024.2839

ABSTRACT

This study involved the development of mathematics achievement test using item response theory. The study was carried out in Rivers State of Nigeria. It was guided by six research questions. The research design adopted in the study was instrumentation research. Mathematics Achievement Test (MAT), developed by the investigators was the instrument for data collection. Table of specifications was used to ensure the content validity The population of this study consisted of all the SS 3 Students in two hundred and seventy-eight (278) public secondary schools in Rivers State, Nigeria, numbering 59,223. The researchers used two different samples of 1, 812 and 2,077 students obtained through a combination of multi-stage and cluster sampling technique. The first sample was used for trial testing of the initial draft of the instrument, while the second sample was used for final testing of the instrument to establish norms and reliability of the final scale. The reliability coefficients obtained using kuder Richardson 20 and Test Re-test techniques were 0.85 and 0.97 respectively, indicating high reliability. The analysis of scores obtained from trial testing of the instrument was done using Mplus software. Research Questions 1, 2 ,3 and 5 were answered using maximum likelihood estimation techniques of Mplus 7.0 software, Research Question 4 were answered using Standard error of measurement while Mean and standard deviation were used to answer Research Question 6.

The results showed that 45 items survived the item analysis in the Mathematics Achievement Test. The means (norms) for female and male' students as measured by the instrument were 23.1650 and 23.2304 respectively. The final instrument is valid and reliable .One of the major recommendations was that workshops and seminars on test development and validation should be organized for classroom teachers and test developers who are not familiar with Item response theory.

Keywords : Item response theory, achievement test , Mathematics , Development

Introduction

Mathematics is one of the essential subjects offered by students and pupils at both primary and secondary school education level in Nigeria. Okigbo, Okeke and Mbakwe (2016) affirmed that among other physical science subjects, mathematics is the backbone in building science and technology. This is as a result of the fact that mathematics equips individual with the ability to enumerate, calculate, measure, collate, group, analyze as well as relate quantities and ideas among others. In Arts and Humanities, mathematical concepts such as measurement, enlargement, symmetry, sequence, proportion, angle of elevation and depression, provide the baseline for better understanding of some related universal concepts like earth and the space (Martin, 2010). Okafor (2015) is of the opinion that, for a nation such as Nigeria to aspire towards scientific and technological development, there is need to pay due attention to mathematics. Mathematical background creates a gap between developed and underdeveloped countries of the world. Mathematics is a language in which scientific ideas are expressed; it is the means by which other sciences as well as Physics, Chemistry, Biology, and disciplines like Engineering, Geology are understood. Mathematics enables the various sciences to draw the implications of their observational and experimental findings. Hence a pass at credit level in the secondary school serves as a pre-requisite for gaining admission into higher institutions.

In spite of the importance and usefulness attached to mathematics, Kurumeh (2006) reported that students achieved poorly in public examinations in mathematics. Also, the chief examiner's reports (2018-2021) from West African Examinations Council (WAEC) show that students' achievement in Senior School Certificate Examination SSCE May/June in Mathematics has not been encouraging. Alshatti (2012) in his study opined that the poor achievement of students in mathematics is orchestrated by ineffective instructional skills and methodologies used by mathematics teachers and students fear for the subject. Other factors which can also contribute to poor achievement in mathematics may include poor learning facilities in schools, inadequate number of trained teachers to teach the subject particularly at the secondary level and more significantly the quality of mathematics achievement test use in assessing the students .

Achievement tests measure the present proficiency, mastery and understanding of general and specific areas of knowledge. According to Nworgu (2015) achievement tests are designed to measure the outcome of the level of accomplishment in a specified programme of instruction in a subject area which

a student had undertaken in recent past. Nworgu explained further that achievement test can be classified, based on quality. into two which are: Teacher-made (classroom) tests and standardized tests. Teacher –made tests are tests constructed and administered by the classroom teacher for the purpose of measuring the attainment of the objectives by the students. Most tests used in our system are of this type. It is assumed that the teacher is in the best position to know the characteristics of his students and also determine the instructional objectives and construct test items of appropriate difficulty to measure the achievement of a certain class of students with peculiar circumstances.

One major limitation of teacher- made tests is that they are narrow in scope(i.e. the items covers a few topics taught within a specified period) and also the psychometric characteristics of the items and tests are usually not established. On the other hand, standardized tests are more carefully and accurately constructed by test experts by adopting elaborate procedures and degree of precision.

According to Okoye (2015), a standardized test is one which was constructed following laid down procedures adopted by measurement experts. Okoye explained that this type of test covers a wide range of objectives and subject matter stipulated in a standard syllabus for a specified level of education.

Standardized tests are not restricted to use in a school or a few schools but to larger population, so that many schools can use such types of tests to assess their own performance.etc. in relation to others and the general population for which the test has been standardized.

The ultimate purpose of a standardized test is, as the name implies, standardization; it provides a standard for comparison. Standardized tests may be designed to evaluate and then compare the aptitudes or competencies of a diverse population of individuals (e.g., students from different institutions who have different educational backgrounds).

Test has different formats which include the following : Essay test and Objective test. Under the objective test , the various types of test include short-answer test , alternative –response items ,matching test items and multiple- choice test items. This study was concerned with multiple-choice test. Multiple-choice test is one in which respondents are asked to select the best possible answer (or answers) out of the choices from a list. The multiple-choice format is most frequently used in educational testing. Multiple-choice test items consist of a stem and a set of options. The stem is the beginning part of the item that presents the item as a problem to be solved, a question asked of the respondent, or an incomplete statement to be completed, as well as any other relevant information. The options are the possible answers that the examinee can choose from, with the correct answer called the key and the incorrect answers called distracters (Akande, 2006). Only one answer can be keyed as correct.

In Psychology and education, measuring instruments are developed based on two major frameworks - the classical test theory and the item response theory. The instrument in this study was therefore developed based on the item - response theory.

Item response theory (IRT) (also known as latent trait theory, strong true score theory, or modern mental test theory) is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. Henard (2000) stated that item response theory is a modeling technique that attempts to describe the examinee's test performance and the latent trait underlying the performance. Similarly, Rivera (2007) opined that item response theory is a body of theory describing the application of mathematical models to data from questionnaire and tests as basis for measuring things such as abilities and attitudes. IRT anchors on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters.

Item response theory is viewed as an improvement over CTT, is more sophisticated and allows for the improvement of the reliability of an assessment. It lays emphasis on three assumptions namely: dimensionality of trait, local independence of items and item response function (IRF) also called Item Characteristic Curve (ICC). Hence latent trait theory has ability to estimate the parameters of an item independent of the characteristics of both the test takers to which it is exposed and other items that make up the test. IRT has the assumption that the examinees' performance can be completely predicted or explained from one or more abilities. It makes provision for more adaptable and effective method of test construction, analysis and scoring than those derived from CTT. Another benefit of item response theory approach in test development is that the parameters of the person do not depend on the parameters of the items, and vice versa. Also in item response theory, the standard error of measurement gives precision at each level of the ability being measured.

Many researchers have carried out studies on development and validation of instruments using item response theory. Uzo (2016) developed and calibrated a basic science achievement test using the two-parameter IRT model. The results of the study showed among others that all the item parameter estimates and person parameter estimates were within the acceptable range, all the items, except one, showed fit to the two parameter IRT model and there was variation in the mean ability estimates of students in the different schools and different local government areas. In the same vein, Ani (2014) applied item response theory in the development and validation of multiple -choice test in Economics. Instrumentation research design was used for the study. The result of the study showed that 49 items of the multiple choice question in Economics were reliable based on three parameter linear (3pl) model. The findings also showed that thirty-one (31) items of the Economics multiple-choice test were difficult. Similarly, Enunwah (2013) investigated the development and standardization of achievement test in SS mathematics using item response theory (IRT) framework. The instrument was suitable for IRT because IRT assumptions of unidimensionality and local independence of items were satisfied by the MCTAPIM data. Item difficulty parameters fell within – 0.55 and 3.13 logit units. The items separated the students into 15.84 strata according to item difficulty parameter. Student ability parameters fell within – 0.79 and 4.37 logit units. Furthermore , Esomonu and Erutujiro (2021) developed and validated Geography Diagnostic Test for senior secondary school students using item response theory. The study adopted instrumentation research design. Based on the analysis, it was found out that: the instrument was unidimensional, the three parameter model of item response theory represented the best fit of the instrument data. Sixty (60) items in the instrument fitted the three parameter model..The empirical reliability of the instrument was 0.98. The final instrument was considered valid and reliable. In another development, Ezechukwu, Oguguo, Ene and Ugorji (2020) developed Economics Achievement

Test (EAT) using Item Response Theory (IRT). Two popular IRT models namely, one-parameter logistic (1PL) and two-parameter logistic (2PL) models were utilized. The researchers adopted instrumentation research design. The finding of the study revealed significant difference between the item difficulties estimated using 1PLM and 2PLM. Also the observed scores of the testees on the test items fited the 1PL 2PL models.

Despite the benefit of item response theory in calibration and standardization of test instruments, most teachers still rely on poorly developed instrument in assessment of students' academic achievement in most subjects including mathematics (Abonyi, 2011). Obinne (2013) and Adedoyin (2010) in their separate studies reported that most achievement tests used by teachers have poor evidence of validity and reliability. The use of poorly developed instruments by teachers on students could yield scores that do not reflect true ability of students in schools. Hence, decision taken based on such scores could be harmful to students and even to school authorities. Therefore, the use of valid and reliable instrument in students' assessment in mathematics should not be underestimated by school authorities and other stakeholders in the education sector.

The scarcity of valid and reliable mathematics achievement test as reflected in the literature could be traceable to teachers' poor knowledge in development of mathematics achievement test or lack of sufficient time on the part of mathematics teachers to develop valid and reliable instruments . In order to cover this gap and challenge, this study was aimed at developing valid and reliable mathematics achievement test that will help ascertain the strength and weaknesses of students in mathematics at senior secondary school level using item response theory .

The aim of this study was to develop Mathematics Achievement Test (MAT) for senior secondary schools using three parameter model of item response theory. Specifically, the study is aimed at achieving the following objectives:

1. Determine the item difficulty indices of the draft Mathematics Achievement Test items
2. Determine the item discrimination indices of the draft Mathematics Achievement Test items.
3. Determine the item guessing indices of the draft Mathematics Achievement Test items.
4. To determine the standard error (empirical reliability) for the draft Mathematics Achievement Test items using IRT
5. To establish the norms for male and female students as measured by the instrument
6. To determine the items that make up MAT.

Considering the specific objectives above, the research questions below guided this study:

1. What are the difficulty indices of the draft Mathematics Achievement Test ?
2. What are the discrimination indices of the draft Mathematics Achievement Test ?
3. What are the item guessing indices of the draft Mathematics Achievement Test?
4. What are the standard errors (empirical reliability) of measurement for the draft Mathematics Achievement Test?
5. What are the norms for male and female students as measured by the instrument?
6. What are the items that make up the Mathematics Achievement Test?

Method

The design of the study was instrumentation research design. Abonyi (2011) defined instrumentation research as a plan of study that enables researchers to develop and often validate instrument required for execution of prescribed tasks. Abonyi further stated that the main purpose of instrumentation research is to create a new assessment facility for educational practice. Therefore, this design was adopted in this study to develop the Mathematics Achievement Test (MAT), which will be used for evaluating the cognitive learning outcomes of senior secondary school mathematics students. The study was carried out in Rivers State of Nigeria.

The population of this study consisted of the SS3 students of the two hundred and seventy-eight (278) public secondary schools in Rivers State, Nigeria, numbering 59,223. (Planning, Research & Statistics (PRS) Department , Rivers State Senior Secondary Schools Board (RSSSB) , 2023). The choice of SS3 mathematics students for the study was as a result of the fact that they had the background knowledge of the subject than any other class in the senior secondary school.

Two different samples were used in this study. The first sample was used for trial testing of the initial draft of the instrument after preliminary validation has been done while the second sample was used for the final testing of the instrument to establish norms and reliability of the final scale . The sample used for the study consisted of one thousand, eight hundred and twelve (1,812) SS3 students. The researchers used a combination of multi-stage and cluster sampling techniques, to obtain the sample.

The instrument was developed based on the procedure for development of mathematics achievement test .The instructional contents and learning outcomes as contained in the National curriculum for senior secondary school mathematics were analyzed and the specific tasks presented by them as implied in each objective were identified. The table of specifications (or test blue print) was then designed in line with the contents and objectives of the mathematics curriculum to guide the researchers in generating the test items. A pool of 120 items was generated each with four response options A-

E. The 120 items spread over the content (topics) and their cognitive levels as contained in the table of specifications. Thereafter, preliminary validation was carried out . Thus the initial instrument was given to two experts in the Department of Educational Foundations, Nnamdi Azikiwe University, Awka and one senior secondary school mathematics teacher. These experts were used to establish content validity of the instrument. A scoring guide was given to each of the validators. At the end of the validation by experts, corrections were made, few questions were modified. None of the items was dropped, because all of them scaled through. These items were trial-tested on SS3 students in the study area with the help of their mathematics teachers in each of the sampled schools. The mathematics teachers that were involved in the trial-testing were guided by the researcher to ensure uniformity in the administration of the draft instrument. The scores obtained from the trial testing of the draft instrument on students were subjected to item analysis in order to ascertain the difficulty, discrimination and guessing indices. Standard errors of each items were estimated as well. The draft instrument was reduced to 45 items which constituted the final instrument..

The 45- item instrument was then administered on a sample of 2077 SS3 students obtained through a combination of multi- stage and cluster sampling techniques in order to ascertain the reliability and norms of the instrument. The reliability of the instrument was established through test-retest technique and kuder Richardson 20. For the test re-test, the final instrument was administered to the SS 3 students. Two weeks later, the same copy of the instrument was re-administered to the same group (sample) to ensure consistency in the reliability coefficient. After the result, a test re-test reliability coefficient of 0.97 was obtained, showing a high reliability of the final instrument. Also, the norms obtained for both male and female students were 25.1650 and 25.2304 respectively.

Research Questions 1, 2, 3 and 5 were answered using maximum likelihood estimation techniques of Mplus 7.0 software. For Item difficulty parameters, any item with index between -2 to +2 was considered and retained (Baker, 2017). For Item discrimination parameter, Baker (2017) interpreted discrimination indices as follows:

0.01 - 0.34 --- Low

0.35 – 1.34 --- Moderate

1.35 – 1.69 --- High

1.70 and above --- Very High

Going further , Baker indicated that any item with discrimination index of 0.35 or above is considered good item and should be retained. Items with guessing value of 0.26 and above were considered to be bad while items with guessing value of 0.25 and below were considered desirable (Harris, 2005). Also , Standard error was used to answer Research Question 4.

Results and Discussion

Research Question 1: What are the difficulty indices of the test items of the draft Mathematics Achievement Test?

Table 1

Difficulty Indices of the Test Items of the Draft Mathematics Test

Item	B	Item	B	Item	B	Item	B	Item	B
1	-1.10	25	2.22	49	3.46	73	0.61	97	1.16
2	-3.98	26	1.04	50	-2.11	74	1.10	98	0.21
3	.410	27	-2.10	51	-.14	75	1.11	99	1.17
4	3.21	28	-1.11	52	1.11	76	3.95	100	1.14
5	-1.13	29	-.24	53	1.14	77	0.83	101	2.07
6	2.21	30	-.11	54	1.72	78	1.89	102	1.10
7	1.41	31	3.18	55	2.20	79	2.18	103	1.31
8	4.61	32	0.21	56	2.12	80	1.75	104	-3.12
9	1.31	33	3.78	57	1.31	81	1.81	105	3.66
10	-2.12	34	1.31	58	-4.11	82	-1.18	106	1.16
11	-3.19	35	2.24	59	2.10	83	-1.09	107	-2.10
12	-1.13	36	3.81	60	-3.76	84	-1.11	108	-1.18
13	2.11	37	2.31	61	-2.11	85	-1.11	109	2.19

14	1.20	38	1.14	62	-1.11	86	-1.81	110	1.16
15	2.31	39	1.16	63	2.14	87	3.11	111	1.17
16	1.21	40	2.20	64	2.14	88	2.16	112	1.07
17	2.14	41	-3.64	65	1.14	89	1.81	113	-3.10
18	1.18	42	1.01	66	-2.21	90	1.17	114	0.14
19	1.10	43	-2.20	67	3.14	91	1.18	115	1.98
20	1.22	44	3.19	68	-2.15	92	2.21	116	1.07
21	1.33	45	1.99	69	2.11	93	2.98	117	0.22
22	2.34	46	0.21	70	1.11	94	2.22	118	1.27
23	2.66	47	1.07	71	2.03	95	2.35	119	1.16
24	1.18	48	2.05	72	2.19	96	2.07	120	0.98

Based on the guideline provided by Baker (2017) for item difficulty parameter, any item between -2 to + 2 is good and should be retained. Table 1 shows that 39 items that is Items 4, 6, 8, 13, 17, 18, 22, 23, 25, 31, 33, 35, 36, 37, 40, 44, 48, 49, 55, 56, 59, 63, 64, 67, 69, 71, 72, 76, 79, 87, 88, 92, 93, 94, 95, 96, 101, 105 and 109 fall above +2 on the logit scale, indicating difficult items while 15 items, that is Items 2, 10, 11, 27, 41, 43, 50, 58, 60, 61, 66, 68, 104, 107 and 113 fall between below -2 on the logit scale indicating easy items. Sixty sixty(66) items, that is Items 1, 3, 5, 7, 9, 12, 14, 16, 19, 20, 21, 24, 26, 28, 29, 30, 32, 34, 38, 39, 42, 45, 46, 47, 51, 52, 53, 54, 57, 62, 65, 70, 73, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 86, 89, 90, 91, 97, 98, 99, 100, 102, 103, 106, 108, 110, 111, 112, 114, 115, 116, 117, 118, 119 and 120 with their indices between -2 to + 2 are good items and were retained.

Research Question 2: What are the discrimination indices of the test item of the draft Mathematics Achievement Test?

Table 2

Discrimination Indices of the Draft Mathematics Achievement Test Items

Item	A	Item	A	Item	A	Item	A	Item	A
1	1.87	25	.37	49	.16	73	.48	97	3.10
2	.12	26	1.88	50	2.10	74	.77	98	.19
3	.14	27	.68	51	3.10	75	1.22	99	.39
4	.15	28	1.92	52	1.24	76	.10	100	.46
5	2.30	29	.28	53	.68	77	.72	101	.16
6	.38	30	1.21	54	.17	78	2.14	102	1.70
7	1.22	31	.24	55	1.14	79	3.18	103	.48
8	.18	32	.94	56	.40	80	.38	104	.10
9	.44	33	.25	57	1.17	81	2.28	105	.11
10	.11	34	.82	58	.18	82	.20	106	2.14
11	.10	35	2.21	59	.73	83	.37	107	2.22
12	.52	36	.10	60	.19	84	.31	108	3.10
13	3.1	37	.81	61	.64	85	2.18	109	.44
14	.51	38	1.17	62	.58	86	.48	110	.14
15	2.10	39	3.21	63	2.27	87	1.89	111	.43
16	.48	40	.72	64	.39	88	.57	112	1.84
17	.20	41	.18	65	1.24	89	.22	113	.15

18	.62	42	3.10	66	.36	90	2.14	114	.39
19	2.11	43	2.20	67	.20	91	.46	115	.48
20	3.10	44	.19	68	2.10	92	.74	116	2.15
21	2.11	45	2.34	69	2.23	93	.74	117	2.34
22	2.23	46	2.22	70	2.11	94	.24	118	.26
23	2.11	47	1.34	71	2.16	95	.77	119	2.78
24	.67	48	2.45	72	2.19	96	2.22	120	2.22

Table 2 shows that 30 items, that is Items 2, 3, 4, 8, 10, 11, 17, 29, 31, 33, 36, 41, 44, 49, 54, 58, 60, 67, 76, 82, 84, 89, 94, 98, 101, 104, 105, 110, 113 and 118 are within the range of .00 to .34 indicating low discriminating power, while 47 items, that is Items 6, 7, 9, 12, 14,16. 18, 24, 25, 27, 30, 32, 34, 37, 38, 40, 47, 52, 53, 55, 56, 57, 59, 61, 62, 64, 65, 66, 69, 73, 74, 75, 77, 80, 83, 86, 88, 91, 92, 93, 95, 99, 100, 103, 109, 111, 114 and 115 are within the range of .35 to 1.34 indicating moderate discriminating power.. Furthermore, 43 items ,that is Items 1, 5, 13. 15, 19, 20, 21,22, 23, 26, 28, 35, 39, 42, 43, 45, 46, 48, 50, 51, 63, 68, 69,70, 71, 72,78,79, 81, 85, 87, 90,96,97.102,107,108,112,116,117,119 and 120 are above 1.70 indicating very high discriminating power. Based on the guidelines by baker (2017), 90 items had acceptable discrimination indices.

Research Question 3: What are the guessing parameters of the test items of the draft mathematics achievement test?

Table 3

Guessing Indices of the Test Items of the Draft Mathematics Achievement Test

Item	C	Item	C	Item	C	Item	C	Item	C
1	.14	25	.28	49	.48	73	.29	97	.29
2	.36	26	.24	50	.18	74	.34	98	.29
3	.05	27	.37	51	.57	75	.24	99	.18
4	.27	28	.25	52	.19	76	.41	100	.11
5	.24	29	.26	53	.29	77	.24	101	.27
6	.52	30	.22	54	.47	78	.18	102	.22
7	.04	31	.31	55	.18	79	.75	103	.21
8	.37	32	.23	56	.04	80	.19	104	.61
9	.05	33	.49	57	.07	81	.16	105	.36
10	.26	34	.24	58	.51	82	.38	106	.18
11	.75	35	.22	59	.08	83	.13	107	.41
12	.07	36	.52	60	.36	84	.36	108	.19
13	.08	37	.27	61	.07	85	.14	109	.18
14	.04	38	.11	62	.08	86	.67	110	.79
15	.74	39	.21	63	.28	87	.04	111	.15
16	.02	40	.15	64	.11	88	.07	112	.14
17	.31	41	.38	65	.13	89	.51	113	.54
18	.25	42	.14	66	.19	90	.08	114	.24
19	.22	43	.18	67	.44	91	.09	115	.22
20	.46	44	.68	68	.27	92	.19	116	.33
21	.01	45	.15	69	.22	93	.23	117	.02
22	.09	46	.18	70	.32	94	.08	118	.04
23	.08	47	.23	71	.34	95	.24	119	.06

24	.17	48	.26	72	.28	96	.04	120	.24

According to the guideline provided by Harris (2005) for selection of items based on guessing parameter, any item with guessing value of .26 and above is not good, while items with guessing value of .25 and below is desirable. Table 3 shows that 48 items had guessing parameters above .25 indicating high probability of guessing. Therefore, Items 2, 4, 6, 8, 10, 11, 15, 17, 20, 25, 27, 29, 31, 33, 36, 37,41, 44, 48, 49, 51, 53, 54, 58, 60, 63, 67, 68, 70, 71, 72, 73, 74, 76, 79, 82, 84, 86, 89, 97, 98, 101, 104, 105, 107, 110, 113 and 116 were rejected due to high probability of examinees guessing them correctly. Table 3 shows that 72 items, that is Items 1, 3, 5, 7, 9, 12, 13, 14, 16, 18, 19, 21, 22, 23, 24, 26, 28, 30, 32, 34, 35, 38, 39, 40, 42, 43, 45, 46, 47, 50, 52, 55, 56, 57, 59, 61, 62, 64, 65, 66, 69, 75, 77, 78, 80, 81, 83, 85, 87, 88, 90, 91, 92, 93, 94, 95, 96, 99, 100, 102, 103, 106, 108, 109, 111, 112, 114, 115, 117, 118, 119, 120 had guessing values less than .26 indicating that they are good items .

Research Question 4: What is the standard error of the mathematics achievement test?

Table 4

The standard error of measurement of the draft mathematics achievement test items

Item	SE	Item	SE	Item	SE	Item	SE	Item	SE
1	.02	25	.34	49	.68	73	.01	97	.01
2	.24	26	.01	50	.00	74	.44	98	.48
3	.02	27	.35	51	.61	75	.01	99	.00
4	.34	28	.02	52	.00	76	.44	100	.02
5	.03	29	.00	53	.01	77	.01	101	.39
6	.67	30	.00	54	.74	78	.00	102	.03
7	.01	31	.47	55	.01	79	.86	103	.00
8	.18	32	.00	56	.04	80	.01	104	.71
9	.00	33	.08	57	.03	81	.04	105	.60
10	.22	34	.01	58	.35	82	.29	106	.01
11	.14	35	.02	59	.02	83	.02	107	.11
12	.03	36	.63	60	.68	84	.37	108	.00
13	.03	37	.00	61	.00	85	.02	109	.01
14	.02	38	.01	62	.01	86	.08	110	.87
15	.61	39	.02	63	.39	87	.04	111	.00
16	.06	40	.00	64	.00	88	.02	112	.03
17	.20	41	.67	65	.00	89	.67	113	.54
18	.00	42	.01	66	.01	90	.00	114	.01
19	.01	43	.02	67	.64	91	.00	115	.02
20	.01	44	.02	68	.00	92	.01	116	.36
21	.03	45	.24	69	.00	93	.02	117	.00
22	.22	46	.00	70	.23	94	.39	118	.04
23	.21	47	.20	71	.22	95	.02	119	.05
24	.04	48	.28	72	.37	96	.03	120	.02

Based on the guideline provided by Obinne (2013), standard error of .05 and below indicates high reliability, while error above .05 indicates low reliability. Table 4 shows that 46 items that is Item 2, 4, 6, 8, 10, 11, 15, 16, 22, 23, 25, 27, 31, 33, 36, 41, 45, 47, 48, 49, 51, 54, 58, 60, 63,67, 70, 71, 72, 74, 76, 79, 82, 84, 86, 89, 94, 98, 101, 104, 105, 105, 107, 110, 113, and 116 had standard error of measurement above .05 and were rejected while 74 items, that is Items 1, 3, 5, 7, 9, 12, 13, 14, 17, 18, 19, 20, 21, 24, 26, 28, 29, 30, 34, 32, 34, 37, 38, 39, 40, 42, 43, 44, 46, 50, 52, 53, 55, 56, 57, 59, 61, 62. 64, 65, 66, 68, 69, 73, 75, 77, 78, 80, 81, 83, 87, 88, 90, 91, 92, 93, 95, 96, 97, 99, 100, 102, 103, 106, 108, 109, 111, 112, 114, 115, 117, 118, 119 and 120 have standard error of .05 and below, indicating high reliability.

Research Question 5: What are the norms for Male and Female students as measured by the instrument?

To obtain the mean for male and female students, the final instrument was administered to a sample of 2,077 students and the scores obtained were used to compute the means and standard deviations for Males and females students' ability in mathematics. The results were presented in Table 5.

Table 5

Mean and Standard Deviation of Male and Female Students as Measured by the Instrument

Gender	N	Mean	SD
Female	1109	25.1650	8.23421
Male	968	25.2304	8.2339

Table 5 shows that the mean achievement score for male students is slightly higher than the mean achievement scores of female students. In other words, male students perform better than female students in the mathematics achievement test.

Similarly, Kuder Richardson 20 and Test Re-test methods of reliability estimation were employed, to ascertain whether the final instrument were reliable and the results is shown in Table 4.10 below.

Table 5.1.

Reliability coefficient of the d Kuder Richardson 20 test re-test techniques

Reliability techniques	R	Remarks
Test Re-test	.97	High
Kuder Richardson 20	.85	High

Table 5.1 showed that the reliability coefficient obtained using Kuder Richardson 20 technique is .85 while the coefficient of reliability using test re-test is 0.97. Hence the, the instrument is reliable.

Research Question 6: What are the items that constitute the final Mathematics Achievement Test?

The number of items that survive the item analysis with their parameters were presented in Table 6 below .

Table 6

The Item Parameters of the Final Mathematics Achievement Test

Item	A	B	C	SE	Item	A	B	C	SE
1	1.87	-1.1	0.14	0.02	65	1.24	1.14	0.13	0.00
5	2.30	-1.13	0.24	0.03	75	1.22	1.11	0.24	0.01
7	1.22	1.41	.004	0.01	77	0.72	0.83	0.24	0.01
9	0.44	1.31	0.05	0.00	78	1.14	1.89	0.18	0.00
12	0.52	-1.13	0.07	0.03	80	0.38	1.75	0.19	0.01
14	0.51	1.20	0.04	0.02	81	2.28	1.81	0.16	0.04
18	0.62	1.18	0.025	0.00	83	0.37	-1.09	0.13	0.02
19	2.11	1.10	0.22	0.01	85	2.18	1.11	0.14	0.02
21	2.11	1.33	0.01	0 .03	90	2.14	1.17	8.08	0.00
24	0.67	1.18	0.17	0.04	91	0.46	1.18	0.09	0.00
26	1.88	1.04	0.24	0.01	99	0.39	1.17	0.18	0.00
28	1.92	-1.11	0.25	0.02	100	0.46	1.14	0.11	0.02
30	1.21	-.11	0.22	0.00	102	1.70	1.10	0.22	0.03

32	0.94	0.21	0.23	0.01	103	0.48	1.31	0.21	0.00
34	0.82	1.31	0.24	0.01	106	2.14	1.16	0.18	0.01
38	1.17	1.14	0.11	0.01	108	3.10	-1.18	0.19	0.00
39	3.21	1.16	0.21	0.02	111	0.43	1.17	0.15	0.00
42	3.10	1.01	0.14	0.01	112	1.54	1.07	0.14	0.00
46	2.22	0.21	0.18	0.00	114	0.39	0.14	0.24	0.01
47	1.34	1.07	0.23	0.02	115	0.48	1.98	0.22	0.02
52	1.24	1.11	0.19	0.00	117	2.34	0.22	0.02	0.00
57	1.17	1.31	0.07	0.03	119	1.79	1.16	0.00	0.05
62	0.58	-1.11	0.08	0.01					

Table 5 shows that 45 items survived the item analysis. Item difficulty indices ranged from -1.89 to 1.99, discrimination indices ranged from 0.37 to 3.21, guessing indices ranged from .01 to .25. while standard error ranged from .00 to .05 .

Discussion

In terms of difficulty, 39 items fell above +2 on the logit scale , indicating difficult items , 15 items fall below -2 on the logit scale indicating easy items while 66 items with their indices between -2 to +2 were retained as good items. The result revealed that all the retained test items were appropriate for measuring examinees of different abilities. A good test item should neither be too difficult nor easy for the examinee. This in line with the suggestion by Dadughan (2015) that a good test item should not be too difficult for examinee, at the same time it should not be too easy for them. Also the finding is similar to the result of the study by Agwagah in Ubada (2000). In the study, 50 items constituted the Mathematics Achievement Test.

Majority of the test items had good discriminating values, as 54 items were within the range of 0.34 to 1.34 , indicating moderate discriminating power, 2 items within the range of 1.35 to 1.66 indicating high discriminating power, 34 items fell above 1.70 indicating very high discriminating power while only 32 items fell between 0.00 to 0.34 indicating low discriminating power . The implication of the above is that most of the items in the instrument can discriminate between high and low achievers in the test.

On the other hand, 73 items had guessing values less than 0.25 indicating that they are good items while 47 items had guessing values above 0.25 indicating high probability of guessing. A good item should not have high guessing parameter as this can make examinee of low ability to score very high. This is in agreement with Young (2014) who recommended that test items of low guessing value should be accepted. The instrument could therefore be used to discriminate between lower and higher achievers .

In a similar development, the study revealed that 76 items had standard error below .05 which indicates high reliability. The standard error of measurement allows researchers to determine the probable range within which the individual's true score falls. The result is similar to that of Obinne (2013) who reported a standard error of .05 and below as implying high reliability, while an error of .05 or less implies low reliability. The instrument is reliable because, according to Meredith et al (2007) , the smaller the standard error of measurement , the more reliable the instrument is . According to Chatterji (2003), standard error of measurement is a statistical estimate of the amount of random error in the assessment of results or scores. This value is similar to the value of reliability coefficients obtained by Adonu (2016), who conducted an intensive study on development and preliminary validation of an instrument for assessment of psychomotor skills in Physics which was found to be 0.87. Onah (2014) conducted a research on development and standardization of Agricultural Science achievement test for senior secondary school students. The reliability value was found to be 0.92. Okereke (2008), found the reliability index for mathematics achievement test to be 0.80. These values of reliability presented above were considered high and therefore the reliability obtained in the present study was also considered to be high. The high reliability index calculated for the present study instrument is not surprising because the instrument was adequately face- and content - validated before administration.

Conclusion

Based on the results of the analysis, forty five (45) items constituted the final form of the mathematics achievement test. Therefore, concluding on the item parameters of the instrument, one would affirm that it is reliable and valid for assessing senior Secondary III students.

Recommendations

As documented throughout the findings, the MAT was developed, validated and scrutinized for empirical evidence of adequacy in measuring the mathematics achievement test. Based on this, it is practically recommended that:

1. Teachers should use the instrument to diagnose persistent learning challenges of students offering mathematics in the senior secondary school so that remedial help can be given or provided.
2. Workshops and seminars on test development and validation using IRT should be organized for classroom teachers and test developers who are not familiar with Item response theory.

ACKNOWLEGEMENT

We appreciate the Principals and students of secondary schools that were used in this study. In the same vein, we also thank the research assistants that assisted in the administration of the instrument to the students in the area of study.

REFERENCES

Abonyi, O. S. (2011). Instrumentation in behavioral research: A practical approach. TIMEX Publishing Company.

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theory. Retrieved from http://www.uniBotswana./journal/ education/science.

Adonu, I.I.(2014).Psychometric analysis of WAEC and NECO practical physics test using partial credit model. Unpublished Ph.D Dissertation. University of Nigeria Nsukka.

Akande, P.F. (2006).The influence of literacy examination system on the development of Chinese civilization, American Journal of Sociology, 35, 336 – 339. of Behavioral and Scale development Research, 2(1), 8-16.

Alshatti, S.A. (2012). Embedding graphic organizers in the teaching and learning of family and consumer sciences in Kuwait. Retrieved from http//eprints. Qut edu.au/6059/1/ safreniz-Ali-Alshatti Thesis (pdfGoole scholar.

Ani, E.N.(2014). Application of item response theory in the development and validation of multiple choice test in economics.Unpublished M.Ed. Thesis. University of Nigeria, Nsukka.

Baker,F.B. (2017). The basics of item response theory. ERIC.

Chatterji, M. (2003).Designing and using tools for educational assessment. Retrieved from http://www.columbia.edu/~mb1434/EdAssess.htm.

Dadughan, S. I.(2015). Development and calibration of primary school mathematics diagnostic test based on item response theory. A Ph. D Thesis submitted to University of Nigeria, Nsukka.

Enunwah, C. I. (2013). Development and standardization of achievement test in SS3 mathematics using Item Response Theory. A Ph.D Dissertation, University of Nigeria, Nsukka.

Esomonu, N. P. M & Erutujiro, G.(2021). Development and validation of geography diagnostic test using item response theory.Journal of Humanities and Social Science, 26(11), 1-9.

Ezechukwu, R. F., Oguguo, B. C. E., Ene, C. U. & Ugorji, C. O. (2020). Psychometric analysis of economics achievement testing using item response theory. World Journal of Education, 10(2), 59-69.

Harris, D. (2005).Educational measurement issues and practice: Comparison of 1-,2-, and 3-parameter IRT models. Retrieved from 10.1111/j.1745-3992. 1989.tb00313.x.

Henard, D.H. (2000), Item response theory, in reading and understanding more - multivariate statistics.American Psychological Association, 2, 67-97.

Kurumeh (2006). Effect of ethnomathmatics approach on students‘ achievement in geometry and menstruation . J. mathematical Association of Nigeria. 31(1) , 35 – 44

Martin, R. (2010). Nigerian societal belief and language effect on the teaching and learning process in science. 32nd Annual Conference Proceeding of STAN 33-36.

Meredith, D. G., Joyce, P. G., and Walter, R.B., (2007).Educational research: An introduction (8th ed.). United States of America: Pearson Press.

Nworgu, B.G. (2015).Educational research: Basic issues and methodology. University Trust Publishers.

Obinne, A.D.E. (2013).Test item validity: Item response theory perspective for Nigeria. Retrieved from www.emergingresource.org.

Okafor, G.B. (2015).Students' perceptions of teaching styles in mathematics learning environments.Journals of Mathematics Teaching Research, 3(2), 1-12.

Okereke, S. C. (2008). Development and preliminary validation of an instrument for the identification of mathematically gifted pupils in Ebonyi State. (Unpublished Ph.D Thesis), University of Nigeria Nsukka.

Okigbo, E. C., Okeke, N.F & Mbakwe, A. B (2016). Enhancing mathematics achievement of introverted and extroverted sec sch students through the use of advance graphic organizers.Educational Research and Review, 4(3), 27-32.

Okoye, R.O. (2015). Educational and psychological measurement and evaluation. Erudition Publishers.

Onah, F. E. (2006). Development and standardization of agricultural science achievement test for senior secondary schools in Enugu State. (Unpublished Ph. D Thesis), University of Nigeria Nsukka.

Rivera, J. E. (2007).Test construction and validation.Developing a statewide assessment for agricultural science.Retrieved from www.spacelibrary.corned.ed/pdf.

Uzo, E.C. (2016). Development and calibration of basic science using 2 parameter logistic model of item response theory.Masters Thesis, University of Nigeria, Nsukka.

West African Examinations Council (2021).Chief examiners' report. Retrieved from waeconline.org.ng/e-learning/mathematics/2.html.

Young, B.A.(2014).“Develoment and validation test in English Language for secondary school in Kisumu Municipality using item response theory“. Unpublished M.Ed. Thesis Univerity of Kassel , Wizehausen , Germany.