IAIEA
International Association For Innovations
In Educational Assessment

JOURNAL OF INNOVATIONS IN EDUCATIONAL ASSESSMENT

Vol. 7, No. 2, June 2025.

Revue des Innovations en Evaluation Pédagogique

ISSN: 2705-3830 (Print)

ISSN: 2705-3857 (Online)

Application of Item Response Theory in the Development and Validation of Multiple-Choice Test in Mathematics

by

Tina Uchenna Otumegwu[1]

Ordua, Victor Ndubuisi.[2]

and

Ifeoma Rosline Ezechukwu [3]

[1,2, & 3]Department of Educational Psychology, School of Education

Federal College of Education (Technical) Omoku Rivers State

otumegwutina@gmail.com O8032718069

Abstract

The study applied three-parameter logistic model (3PLM) of IRT in the development and validation of Mathematics Multiple-Choice Test items. Two research questions were raised along with two hypotheses postulated to guide the conduct of the study. The study adopted instrumentation research design which involved the development of instrument for educational purpose. This study was conducted in Owerri Education Zone of Imo State. A sample of 1080 SS2 students was drawn from a population of 6,823 students that made up of all the SS2 students using simple random sampling techniques. The instrument for data collection for this study is Mathematics Multiple – Choice Test (MMT) developed by the researchers. The reliability coefficient 0.97 for the instrument was computed using Kuder Richardson formula twenty (KR_{20}). Research questions were answered using standard errors of estimation and Chi-square goodness of fit. While hypotheses were tested using p-value, z-test of significant of proportion and z-value for testing the fit of the items to the three-parameter logistic model (3PLM). Some of the findings of the study revealed that forty-seven out of the fifty items have Standard Errors of Estimation (SEE) of discrimination parameters below 0.1, while only three has SEEs of a above the criterion of 0.1. Forty-four items have SEEs of difficulty parameters that are below 0.1, while only six have SEEs of b above 0.1. The study therefore, recommends that secondary school Mathematics teachers should confidently use the items of the test in assessing their students since the items have good standard errors of estimation for the three items parameters.

Keywords: *Item Response Theory (IRT), Multiple-Choice Test, Mathematics and Standard Error of Measurement.*

Application of Item Response Theory in the Development and Validation of Multiple-Choice Test in Mathematics

Mathematics is one of the core subjects in the Senior Secondary School that every student must offer; it also requires assessment to ascertain students' basic knowledge, skills and understanding of the concepts and nature of Mathematics problems, (Otumegwu, Ezechukwu & Ojedapo,2024). According to Srivastav in (Ogbonna and Oluwene, 2018), Mathematics is a fundamental science which deals with the study of time, space, measurement, quantities, shapes and numbers and how they relate with each other. This means that Mathematics cut across almost all aspect of lives. It is a powerful tool for reasoning and problem-solving that influences nearly every aspect of our lives, from the simplest tasks to the most complex scientific theories. Iweka (2017), states that Mathematics is an efficient tool used in all sciences and for technological development of any nation. It is a fundamental part of human knowledge and one of the central planks of the modern technological revolution, (Ernest, 2015). Mathematics provides us with a broad range of skills in problem solving, logical reasoning and flexible thinking, (Jayanthi, 2014).

The Federal Republic of Nigeria (FRN) (2013) in her National Policy on Education states, “Mathematics should be visualized as the vehicle to train a child to think, reason, analyze and to articulate logically”. Mathematics is a fundamental part of human thought and logic, and integral to attempts at understanding the world and ourselves. It provides an effective way of building mental discipline and encourages logical reasoning and mental rigor. Mathematic knowledge plays a crucial role in understanding the contents of other school subjects such as science, social studies, and even music and art.

Mathematics holds an important and unique place among other subjects. The historical role of Mathematics supports the notion that Mathematics has provided the mental discipline required for other disciplines. Mathematics is applied in various fields and disciplines, i.e. mathematical concepts and procedures are used to solve problems in sciences, and social sciences. (For example, the understanding of complex numbers is a prerequisite to learn many concepts in electronics). The complexity of those problems often requires relatively sophisticated mathematical concepts and procedures.

According to Otumegwu, Ezechukwu and Ojedapo (2024), in education, the importance and the place of a particular subject depends on the extent the subject is helpful in achieving the aims and objectives of the education. If any subject is more useful for achieving educational objectives, then its importance increases accordingly. Since ancient times, Mathematics has played a vital role in achieving aims and objectives of education, as compared to others. Present age is the era of science and information technology. Whatever, technological and physical progress being made, shall be correspondent to the role of Mathematics. To achieve the objectives of teaching and learning of Mathematics, different assessment instruments are used such as; essay tests and objective tests which are utilized by the teacher depending on the aims of the measurement.

Ajuonuma (2016) posits that essay tests are tests that permit the students to express their responses in their own words and in the way they deem fit. In essay test, students are offered the opportunity to organize and express their ideas clearly in writing. Objective test is one of the assessment instruments used in testing or assessing students' academic achievement in any given instruction. In objective tests, such as multiple-choice questions, students or respondents are asked or required to select the best possible answer out of the options from the list. Multiple-Choice Test items consist of a direct question or an incomplete statement, which is answered or completed by selecting the answer from the options or a list of suggested solutions, from which one of them is the correct answer. The problem may be a direct question or an incomplete sentence called options or alternatives (Ekwonye, 2015). The author maintained that, the part of the question that bears the task of the item is called the stem and the suggested solutions or listed responses are called alternatives or options. The correct option is known as the Key while the other options which are incorrect are known as distractors/distracters. The testee is expected to read the stem and select the correct or best option as the answer. The distracters are to distract the students who are in doubt about the correct answer. The number of options used in the multiple – choice items ranges from three to five. The form of stem (direct question, or incomplete statement) depends on several factors. The factors include the ease in writing the stem, the level of the testee, clarity in problem formulation, etc. The scores obtained from the multiple-choice questions are used to assess the competence of the students. Some of the advantages of the multiple-choice question tests are; the items can be widely applied to measure simple and complex behaviours (the lower and higher levels of cognition) (Ajuonuma, 2016).

Multiple choice tests are test which allows the teacher and examiners to ask a large number of questions that mostly cover the subject or course content. They are easy to score and can be scored by non-specialists like clerks, students themselves and also by machines. They have very high scorer - reliability and test-retest reliability. They are relatively free from response sets. That is, students generally do not have the tendency to favour a particular alternative when they do not know the answer. However, all assessment instruments must satisfy the criteria of validity, reliability as well as usability. (Anene & Ndubuisi, 2015).

Validity of a test is the extent to which the test measures what it is supposed to measure (Ejimaji & Ojedapo, 2017). It is the degree of accuracy with which a test measures what it is sets out to measure. A valid instrument must measure accurately and consistently what it is designed to measure. A valid test must satisfy the psychometric functions of evaluation (Ukozor, 2016). It means that the test should be able to measure achievement of learners, discriminate them according to their demonstrated abilities and the same time, be appropriate for predicting subsequent outcomes. Reliability is conceived in relation to the extent of the consistency or dependability of a measuring instrument (Abonyi, 2011). This implies that if any test were to be applied in Mathematics several times, it would be expected to generate responses that vary a little from trial to trial, as a result of measurement error. Therefore, for any measuring instrument, the smaller the error, the greater the reliability while the greater the error, the smaller the reliability. Individual scores on a test can be viewed as the combined result of the true score and the measurement error.

The type of measurement error that is utilized in interpreting individual scores is called standard error of measurement. According to Musselwhite and Wesolowski (2018), Standard error of measurement can be used to estimate a range of scores around a specified cut off point when determining an examinee's ability or potential. The normal distribution can aid in the interpretation of scores that fall above, below, or between specific points on the distribution. It provides the standard deviation of a series of measurements taken on the same individual.

Furthermore, instrument developed in Mathematics requires more determination of validity, reliability, objectivity and usability of items. Any test and indeed any evaluation instrument must satisfy the criteria of reliability, validity as well as objectivity (Nworgu, 2015). In the development of a test therefore, a number of steps are involved. These are:

i. Content analysis

ii. Review of the instructional objectives:

iii. Development of test blueprint / Table of specification:

iv. Item writing:

v. Validation

vi. Item review

vii. Trial Testing:

viii. Item analysis

(Anene & Ndubuisi, 2015)

This is what item analysis dwell on. For educational objective to be attained through testing, teachers must comply with constructing tests that are adequate, consistently and conveniently measuring the expected behaviour of the learners. These need to be ensured to build in credibility in the test so that decisions made through their results can be credible.

The procedures for determining these indices or parameter of items of the instrument depend on the measurement theory used. The two distinct measurement theories are the Classical Test Theory (CTT) and Item Response Theory (IRT). Classical test theory (CTT) is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. It is a theory of testing based on the idea that a person's observed or obtained score on a test is the sum of a true score (error-free score) and an error score (Allen & Yen, 2002). Symbolically, $X = X_t + X_e$, where X = any raw score or unit of measurement,

X_t = true score component, $X_{e\,=}$ error score component. Classical test theory is based on the true score theory which views the observed score (X) as a combination of the true scores (T) and an error component (E) (Adedoyin, 2010). The observed score of a test-taker is usually seen as an estimate of the true scores of the test-taker plus or minus some unobservable measurement error (Crocker &Algina, 2018).

Classical Test Theory (CTT) is precisely concerned with the relationship between the variables: observed / raw score of the attribute possessed by an examinees being measured in the time of measurement and the error score (E) which indicates the effects of extraneous influences of the measurement possessed at the time of measurement, and it is considered to be random (Ekwonye & Eguzo, 2011). CTT does not have a complex theoretical model to relate an examinee's ability to succeed on a particular item, but collectively considers a pool of examinees and empirically examinees' ability to success on a particular item, (Ani, 2014). CTT has its shortcomings or disadvantages; examinees' characteristics and test's characteristics cannot be separated: each can only be interpreted in the context of the other. In CTT, the standard error of measurement is assumed to be the same for all examinees. CTT is test oriented, rather than item oriented. In other words, CTT cannot be used to predict how well an individual or a group of examinees might perform on a test item.

Though, despite the limitation of CTT it is being used to estimate the achievement test in the secondary schools. For instance, the students' achievements in Mathematics are often subjected to statistical measurement as mean, standard deviation, etc. These statistics change for a test when another sample from the same population of students is used. The estimates are obtained depending on how many samples chosen from the students' population. This means, it depends on students' aggregate score in a test while the achievement on individual items is not determined.

Therefore, to achieve effective teaching and learning of Mathematics in schools, an achievement test that focuses on the achievement of individual items will be better than the one on students' aggregate scores. Considering the inadequacies and shortcomings observed in the use of Classical Test Theory, scholars developed another test theory called Item Response Theory. This one uses a Mathematics model. An educational measurement scale that has ratio scale, sample independent and students' ability reported on both item and total instrument levels can be developed with the measurement theory called Item Response Theory (IRT) also, known as modern theory. Some researchers assert that Item Response Theory (IRT) is solution for the limitations of classical test theory. Ekwonye & Eguzo, (2011) defined IRT as a theory that studies the test item and response scores of test items based on assumptions concerning the mathematical relationship between abilities and the probability of getting an item right or correct. According to Henard (2000), Item Response Theory is a modeling technique that tries to describe the relationship between an examinee's test performance and the latent trait underlying the performance. Reeve (2002), describes Item Response Theory as a body of theory describing the application of Mathematics models to data from questionnaires and tests as a basis for measuring things such as abilities and attitudes. Item Response Theory (IRT) looks at the examinee's performance by using item distributions based on the examinee's probability of success on a latent variable. In IRT, item statistics also referred as parameters are estimated and interpreted. Under IRT, parameters of the persons are invariant across items, and parameters of the items are invariant in different populations of persons. It brings greater flexibility and provides more sophisticated information which allows for the improvement of the reliability of an assessment. Item Response Theory is a collection of different models showing the relationship between a participant's responses on an item and underlying latent trait (Ercikan& Koh, 2015).

IRT model assumes that the performance of an examinee can be completely predicted or explained from one or more abilities. IRT models the probability of a correct answer using three logistic functions. The One-Parameter Logistic Model (1PLM) (also known as Rash model), adjust the item difficulty level as trait level required for correctly answering a question. It attempts to address the probability of a correct answer by allowing each question to have an independent difficulty variable. For instance, one-parameter model allows each question on an achievement test to have an independent difficulty variable. The Two-Parameter Logistic Model (2PLM), accounts for item difficulty and discrimination parameters. It attempts to model each item's level of discrimination between high and low ability students. While Three-Parameter Logistic Model (3PLM) adds a third item parameter which is called pseudo-guessing parameter that reflects the probability that an examinee with a very low trait level will correctly answer an item in an achievement test by guessing. This model takes into account the effect of item guessing in addition to the difficulty and discrimination level of the item. Also, assumes that the three parameters, difficulty, discrimination and guessing are combined for an estimate of a relationship between the probability of a correct response of an item and the trait level (ability) of an examinee. Obinne (2012) observed that the guessing is giving an answer or making a judgment about something without being sure of all the facts. Guessing parameter model gives the probability of an individual with low ability, answering correctly to an item with a difficulty, discrimination, and guessing index.

In the latent trait test model, the internal validity of a test is assessed in terms of the statistical fit of each item to the model. Fit to the model, implies that item discriminations are uniform and substantial, there are no errors in item scoring. It also indicates that guessing has a negligible effect on test scores. IRT models are very important in assessment instrument like Mathematics achievement test when trying to understand students' abilities by examining their test performance. To ensure that Mathematics achievement test is fair for all examinees, the instrument should be fair. A test instrument is said to be fair for all when two groups of equal ability with respect to the construct measured by the test should get the same score on each item of the test.

Statement of the Problem

Based on the limitations of the instrument developed under classical test theory, the researchers designed this study using a modern measurement theory (IRT) to develop instrument that will ensure objectivity in measurement of the students' achievement in Mathematics. Therefore, the question to be addressed is 'would the application of Item Response Theory, enhance the development of Multiple-Choice Test instrument in Mathematics?'

The main purpose of this study is to apply Item Response Theory in the development and validation of the Multiple-Choice Test in Mathematics. Specifically, the study will determine the;

1. standard errors of measurement of the items of the Multiple-Choice Test in Mathematics.
2. overall model fit of the Mathematics Multiple Choice Test using Three-Parameter Logistic (3PL) model.

Research Questions

1. What are the standard errors of estimation of the item parameters of the Multiple-Choice Test in Mathematics developed using IRT?
2. How do the items of the Multiple-Choice Test fit the Three-Parameter Logistic Model (3PLM)?

Hypotheses

1. The proportion of items of the Multiple-Choice Test in Mathematics that have standard errors of estimation of the item difficulty parameters that are within the acceptable value is not significantly greater than 0.5.
2. The items of Mathematics multiple choice test do not significantly fit the Three-Parameter Logistic Model (3PLM).

Method

The instrumentation research design was adopted in this study. The target population was made up of all the SS2 students in all the public secondary schools in Owerri Education Zone of Imo State, total of 6823 students and 119 schools. A sample of 1080 SS2 students from 54 public co-education secondary schools was drawn using random sampling technique. The instrument for data collection for this study is Mathematics Multiple – Choice Test (MMT) developed by the researchers. The (MMT) was constructed based on the topics drawn from SS2 syllabus. The instrument consisted of 50 multiple choice questions, each with 4 options (A – D). One mark was assigned to each correct response and zero for incorrect response. Scoring guide which contained all the answers to the fifty (50) multiple choice questions was developed also by the researchers. MMT was validated by two experienced Mathematics teachers and three specialists in educational measurement and evaluation. Kuder Richardson ($K – R_{20}$) formula was used to determine the internal consistency of the instrument and the reliability estimates for the instrument was 0.97. The data for this study was collected through the use of Mathematics Multiple – Choice Test (MMT), which the researchers' administered on the S2 students of the sampled schools. Research questions were answered using standard errors of estimation and Chi-square goodness of fit. While hypotheses were tested using p-value, z-test of significant of proportion and z-value for testing the fit of the items to the three parameter logistic model (3PLM). Decision rule; the criterion for SEE of a- and b-parameters = 0.1, c-parameter =0.05, If observed p-value < 0.05 = misfit, p-value> 0.05 = fit.

Results

Research Question One: What are the standard errors of estimation of the items of the Multiple –Choice Test in Mathematics?

Table 1

Standard Errors of Estimation of the Items of MMT

Item ID	*a*	*b*	*c*	a SE	b SE	c SE	Chi-sq	Df	P	Flags
1	2.638	1.721	0.241	0.096	0.079	0.033	15.9993	12	0.1913	
2	0.982	0.598	0.245	0.096	0.084	0.042	10.9314	12	0.5348	
3	0.624	0.782	0.362	0.063	0.086	0.031	11.7527	12	0.4657	
4	0.524	1.743	0.254	0.57*	0.019	0.024	15.2172	12	0.2298	K La
5	1.263	-1.948	0.252	0.098	0.007	0.023	21.2491	12	0.0468	
6	5.502	1.487	0.25	0.064	0.048	0.026	15.3955	12	0.2205	
7	0.428	0.888	0.249	0.056	0.082	0.036	15.3933	12	0.2206	
8	1.122	1.066	0.25	0.019	0.056	0.042	6.7661	12	0.8727	
9	2.513	1.180	0.252	0.069	0.034	0.036	12.2351	12	0.427	
10	0.362	-1.372	0.249	0.094	0.02	0.045	9.9573	12	0.6197	
11	0.233	-1.324	0.249	0.099	0.03	0.046	15.6692	12	0.2069	
12	0.666	0.854	0.248	0.065	0.069	0.02	14.18	12	0.2894	
13	0.958	1.411	0.252	0.001	0.025	0.046	10.4331	12	0.578	
14	0.406	0.799	0.252	0.2*	0.07	0.049	17.4478	12	0.1335	La
15	1.298	1.642	0.249	0.031	0.016	0.035	17.7021	12	0.125	
16	0.152	3.012	0.252	0.022	0.1*	0.025	19.0006	12	0.0885	K Hb
17	0.28	2.447	0.254	0.049	0.023	0.026	19.3227	12	0.081	K

18	1.491	0.794	0.25	0.085	0.076	0.044	10.8657	12	0.5405	
19	5.502	-0.082	0.249	0.009	0.5*	0.69*	7.1995	12	0.8442	
20	0.308	1.682	0.248	0.008	0.02	0.036	18.1407	12	0.1115	
21	0.598	-0.628	0.248	0.099	0.092	0.5*	15.9283	12	0.1945	
22	0.354	1.540	0.251	0.062	0.016	0.036	8.4315	12	0.7506	
23	4.033	-2.284	0.251	0.049	0.017	0.036	21.9804	12	0.0377	
24	0.256	2.728	0.25	0.025	0.009	0.033	18.5634	12	0.0996	
25	5.502	-0.544	0.249	0.007	0.086	0.5*	13.5998	12	0.327	
26	0.559	3.280	0.249	0.024	0.24*	0.019	18.3794	12	0.1046	Hb
27	5.502	1.757	0.253	0.031	0.036	0.035	11.2675	12	0.5061	K
28	2.376	2.914	0.254	0.025	0.019	0.01	13.9197	12	0.3059	K
29	0.85	1.214	0.25	0.043	0.075	0.047	16.668	12	0.1625	
30	5.502	1.603	0.249	0.041	0.032	0.049	18.5761	12	0.0993	
31	0.181	2.620	0.254	0.035	0.006	0.024	15.649	12	0.2078	K
32	3.11	-0.242	0.251	0.008	0.68*	0.71*	14.209	12	0.2876	
33	0.226	1.474	0.253	0.084	0.084	0.041	22.1777	12	0.0356	
34	2.975	-2.316	0.251	0.046	0.055	0.034	7.4122	12	0.8292	
35	2.264	1.370	0.251	0.016	0.085	0.043	10.4114	12	0.5799	
36	2.317	1.392	0.247	0.065	0.077	0.046	13.6766	12	0.3218	
37	0.215	-2.627	0.247	0.024	0.019	0.01	10.3281	12	0.5872	
38	1.116	-2.529	0.253	0.041	0.036	0.01	16.3254	12	0.1768	K
39	0.389	-2.917	0.251	0.023	0.009	0.013	9.5052	12	0.6593	K
40	3.412	0.973	0.25	0.032	0.71*	0.55*	21.9831	12	0.0377	
41	0.699	2.619	0.251	0.031	0.004	0.01	10.2782	12	0.5916	
42	1.067	-1.461	0.248	0.056	0.061	0.044	11.0484	12	0.5248	
43	2.522	1.566	0.251	0.061	0.039	0.045	19.5317	12	0.0765	
44	1.54	0.711	0.251	0.034	0.077	0.5*	14.4281	12	0.2742	
45	0.991	0.383	0.25	0.013	0.67*	0.54*	9.193	12	0.6864	
46	0.536	0.868	0.253	0.88*	0.087	0.047	15.5224	12	0.2141	K La
47	0.455	2.820	0.251	0.024	0.001	0.027	9.4601	12	0.6632	
48	0.311	1.136	0.251	0.07	0.078	0.048	19.1336	12	0.0854	
49	0.339	-1.589	0.249	0.031	0.046	0.045	12.0562	12	0.4412	
50	0.351	1.697	0.248	0.052	0.036	0.043	13.4332	12	0.3384	

Table 2

Standard Errors of Estimation of a-, b- and c-parameters of Items of MMT

$a < 0.1$	$a \geq 0.1$	$\mathbf{b} < 0.1$	$\mathbf{b} \geq 0.1$	$\mathbf{c} < 0.05$	$\mathbf{c} \geq \mathbf{0.05}$
47	3	44	6	43	7

Table 1 shows the a-, b-, and c-parameters of the 50 items of the developed Mathematics multiple – choice test and their corresponding standard errors of estimation. The items with SEE of a- and b-parameters and c-parameters that are below the SEE criterion of 0.1 and 0.05 are marked (*). As shown in Table 2; 47 items have SEEs of discrimination parameters below 0.1, while only three (items 4, 14, and 46) has SEEs of a above the criterion of 0.1. Similarly, 44 items have SEEs of difficulty parameters that are below 0.1, while only six (items 16, 19, 26, 32, 40, and 45) have SEEs of b above 0.1. On the other hand, 43 items have SEEs of guessing (c) parameters that are below the criterion of 0.05, while only seven (items 19, 21, 25, 32, 40, 44, and 45) items have SEEs above 0.05.

Research Question Two: How do the items of the Mathematics multiple choice test fit the three-parameter logistic model (3PLM)?

Table 3

Item-by-Item Chi-square and z-Resid Goodness of Fit Test

	θ -max	a	b	c	Chi-sq	df	P	z Resid
1	2.24	2.638	1.721	0.241	15.9993	12	0.1913	0.3103
2	1.14	0.982	0.598	0.245	10.9314	12	0.5348	0.2885
3	1.46	0.624	0.782	0.362	11.7527	12	0.4657	0.1071
4	2.4	0.524	1.743	0.254	15.2172	12	0.2298	0.5134
5	2.54	1.263	-1.948	0.252	21.2491	12	0.046*	0.3939
6	2.04	5.502	1.487	0.25	15.3955	12	0.2205	0.1474
7	1.42	0.428	0.888	0.249	15.3933	12	0.2206	0.2156
8	1.64	1.122	1.066	0.25	6.7661	12	0.8727	0.1142
9	1.78	2.513	1.180	0.252	12.2351	12	0.427	0.3382
10	1.94	0.362	-1.372	0.249	9.9573	12	0.6197	0.1777
11	1.88	0.233	-1.324	0.249	15.6692	12	0.2069	0.1462
12	1.38	0.666	0.854	0.248	14.18	12	0.2894	0.2576
13	2	0.958	1.411	0.252	10.4331	12	0.578	0.2512
14	1.42	0.406	0.799	0.252	17.4478	12	0.1335	0.3602
15	2.2	1.298	1.642	0.249	17.7021	12	0.125	0.2774
16	3.58	0.152	3.012	0.252	19.0006	12	0.0885	0.4836
17	3.06	0.28	2.447	0.254	19.3227	12	0.081	0.5496
18	1.36	1.491	0.794	0.25	10.8657	12	0.5405	0.0726
19	0.46	5.502	-0.082	0.249	7.1995	12	0.8442	0.3558
20	2.22	0.308	1.682	0.248	18.1407	12	0.1115	0.2154
21	1.14	0.598	-0.628	0.248	15.9283	12	0.1945	0.2133
22	2.12	0.354	1.540	0.251	8.4315	12	0.7506	0.2158
23	2.84	4.033	-2.284	0.251	21.9804	12	0.037*	0.2277
24	3.28	0.256	2.728	0.25	18.5634	12	0.0996	0.2605
25	1.1	5.502	-0.544	0.249	13.5998	12	0.327	0.3093
26	3.82	0.559	3.280	0.249	18.3794	12	0.1046	0.263
27	2.36	5.502	1.757	0.253	11.2675	12	0.5061	0.4034
28	3.5	2.376	2.914	0.254	13.9197	12	0.3059	0.4768
29	1.78	0.85	1.214	0.25	16.668	12	0.1625	0.074
30	2.16	5.502	1.603	0.249	18.5761	12	0.0993	0.298
31	3.2	0.181	2.620	0.254	15.649	12	0.2078	0.4081
32	0.82	3.11	-0.242	0.251	14.209	12	0.2876	0.136
33	2.06	0.226	1.474	0.253	22.1777	12	0.035*	0.3502
34	2.88	2.975	-2.316	0.251	7.4122	12	0.8292	0.2628
35	1.96	2.264	1.370	0.251	10.4114	12	0.5799	0.3309
36	1.92	2.317	1.392	0.247	13.6766	12	0.3218	0.3734
37	3.14	0.215	-2.627	0.247	10.3281	12	0.5872	0.1837
38	3.12	1.116	-2.529	0.253	16.3254	12	0.1768	0.5429
39	3.48	0.389	-2.917	0.251	9.5052	12	0.6593	0.3346
40	1.52	3.412	0.973	0.25	21.9831	12	0.037*	0.1883
41	3.18	0.699	2.619	0.251	10.2782	12	0.5916	0.314
42	2	1.067	-1.461	0.248	11.0484	12	0.5248	0.1959
43	2.16	2.522	1.566	0.251	19.5317	12	0.0765	0.2993
44	1.3	1.54	0.711	0.251	14.4281	12	0.2742	0.3009
45	0.94	0.991	0.383	0.25	9.193	12	0.6864	0.1797
46	1.52	0.536	0.868	0.253	15.5224	12	0.2141	0.5274
47	3.36	0.455	2.820	0.251	9.4601	12	0.6632	0.2982
48	1.72	0.311	1.136	0.251	19.1336	12	0.0854	0.2438
49	2.12	0.339	-1.589	0.249	12.0562	12	0.4412	0.1433
50	2.22	0.351	1.697	0.248	13.4332	12	0.3384	0.3039

Significant

Table 3 shows the Chi-square, p-value and z for testing the fit of the items to the three parameter logistic model (3PLM). The Chi-square values for the items ranges from 6.766 to 22.178, while the p-value and z values ranges from 0.0356 to 0.873 and 0.073 to 0.550 respectively. An item's calculated Chi-square value greater than the tabulated Chi-square value of 21.026 and the p-value less than 0.05 indicate significant. Based on this, only four (4) items (item 5, 23, 33, and 40) representing 8% of the 50 items did not fit the three parameter logistic model (3PLM). These items (items 5, 23, 33. And 40) are indicated by asterisk (*), with p-values 0.046, 0.037, 0.035, and 0.037. This means that the remaining 46 items (representing 92% of all the items) fitted the three parameter logistic model (3PLM).

Hypothesis One: The proportion of items of the multiple choice test in Mathematics that have standard errors of estimation of the item difficulty parameters that are within the acceptable value is not significantly greater than 0.5.

Table 4

Summary z-test Statistics for Testing Hypothesis One

n	P	z_{Cal}	df	z_{Crit}	$Decision$
50	0.88	8.268	49	1.671	fail to accept H_{01}

From Table 3 the calculated z-test statistic 8.268 is greater than the critical z-test statistic 1.671 at 0.05 level of significance and degree of freedom 49. The researchers, therefore, fail to accept hypothesis one which states that "the proportion of items of the Multiple-Choice Test in Mathematics that have standard errors of estimation of the item difficulty parameters that are within the acceptable value is not significantly greater than 0.5". Hence, the proportion of items of the Multiple-Choice Test in Mathematics that have standard errors of estimation of the item difficulty parameters that are within the acceptable value is significantly greater than 0.5.

Hypothesis Two: The items of Mathematics multiple choice test do not significantly fit the three-parameter model (3PLM).

To test the hypothesis, the p-value and z-value were used. The observed p-value ranges from 0.0356 to 0.873, while the observed z values ranges from 0.073 to 0.550. The items with p-value less than 0.05 indicate significant. These items (items 5, 23, 33, and 40) are indicated by asterisk (*), with p-values 0.046, 0.037, 0.035, and 0.037, indicating that they did not significantly fit the three parameter logistic model (3PLM). This means that the remaining 46 items (representing 92% of all the items) fitted the three parameter logistic model (3PLM).

Discussion

The finding of the study revealed that forty-seven out of the fifty items have standard errors of estimation (SEE) of discrimination parameters below 0.1, while only three has SEEs of a above the criterion of 0.1. Forty-four items have SEEs of difficulty parameters that are below 0.1, while only six have SEEs of b above 0.1. On the other hand, forty-three items have SEEs of guessing (c) parameters that are below the criterion of 0.05, while only seven items have SEEs above 0.05. The standard error of estimation provides a more direct measure of the accuracy of prediction or estimation of the item parameters (a-, b-, and c-parameters). The standard error of estimation for an item parameter indexes the imprecision in the parameter estimates. This shows that majority of the parameters estimated for the items are reliable. The SEEs are used to construct a confidence interval for a given item parameter to help take into account calibration error. Hence, the small standard errors estimation of discrimination parameters for items 4, 14 and 46 indicates the inconsistency of the items in estimating the discrimination parameters. The same applies to the large standard errors of estimation of difficulty parameters for items 16, 19, 26, 32, 40, and 45, and guessing parameters of items 19, 21, 25, 32, 40, 44, and 45. For these items the reliability of their estimation is low since their standard errors of estimation are higher. The standard errors of estimation of these items are higher than the criterion of 0.1. This result agrees with Obinne (2018) and Ani (2014) that SEEs below 0.10 is described as high reliability, while SEEs above 0.10 is described as low reliability.

The finding of the study revealed that four items out of the 50 items of the Mathematics Multiple-Choice Test have statistically significant fit. This indicates that these items did not fit the three-parameter logistic model (3PLM). There is, therefore, misfit for items 5, 23, 33, and 40. However, the remaining 46 items (representing 92% of all the 50 items) fitted the three parameter logistic model (3PLM). These four items that have substantial misfit would be removed from the test and the remainder of the 46 items recalibrated. The criterion for all the item fit/misfit in this study was determined at 0.05 level of significance. This is in agreement with Guyer and Thompson (2014) who proposed that items that have substantial misfit should be removed from a test and the remainder of the items recalibrated. In a similar study Ani (2014) and Adedoyin (2010) determined fit/misfit at 0.05 level of significance and found that majority of their items significantly fit the three-parameter logistic model (3PLM) they used.

Conclusion

The study demonstrated that the majority of the Mathematics Multiple-Choice Test items exhibited reliable estimations for discrimination, difficulty, and guessing parameters. Standard Errors of Estimation (SEEs) below established criteria indicate high reliability for most items, affirming the precision of their parameter estimates. However, items with SEEs exceeding the criteria, such as items 4, 14, 46 (discrimination), 16, 19, 26, 32, 40, 45 (difficulty), and 19, 21, 25, 32, 40, 44, 45 (guessing), were deemed less reliable in their estimations. Additionally, four items – 5, 23, 33, and 40 – did not fit the Three – Parameter Logistic Model (3PLM), signifying misfit and necessitating their removal from the test for recalibration of the remaining items. Overall, 92% of the items fitted the 3PLM, reflecting consistency and alignment with prior research on item reliability and model fit criteria. These findings emphasize the importance of addressing misfit and large SEEs to improve the validity and reliability of test items.

Recommendations

Based on the findings of the study, the following recommendations were made.

1. SSII Mathematics teachers should confidently use the items of the test in assessing their students since the items have good standard errors of estimation for the three items parameters.
2. The items that did not fit the three-parameter logistic model (3PLM) should be recalibrated using 1PLM or 2PLM of IRT.

3. The researchers recommend that the Mathematics Multiple-Choice Test be used extensively in secondary schools by Mathematics teachers during assessment of students, since it contains items with large discrimination parameters and as a result is able to distinguish between students at the low and high ability levels.

4. The Mathematics teachers in secondary schools should use the items of the Mathematics Multiple-Choice Test without worrying so much about students guessing their answers.

References

Abonyi, O.S. (2011). *Instrument in behavioural 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 theories. *International Journal Education Science, 2(2)*, 107 – 113.

Ajuonuma, J. (2016). C*onstruction and administration of teacher – made tests and standardized tests inOgomaka, P.M.C, Ekwonye, E.C, Ukozor, F. I, and Onah, F. E, (*Eds) *measurement and evaluation: A comprehensive text for students and teachers.* Flash point publishers. *Pp218 – 255*

Ali, A. (2014). *Conducting research in education and the social sciences.* Tashiwa Networks Ltd.

Allen, M.J. & Yen, W.M, (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland press.

Anene G.U, & Ndubuisi O.G, (2015). *Test development process. In B.G. Nworgu (Ed) Educational measurement and evaluation: Theory and practice (pp 110 – 122).* University trust publishers.

Crocker, L, & Algina, J. (2018). *Introduction to classical and modern test theory.* Fort Worth: Harcourt Brace Jovanovich.

Ejimaji, E.U., & Ojedapo, D.O., (2017). *Fundamentals of educationa measurement and evaluation*. Jef-printing and publishing Co.

Ekwonye, E.C. (2015). Construction and administration of teacher made and standardized test. In Ekwonye, E.C, Uzoma, P.N, Offor, L, &Eguzo, G.O. *Introduction to testing and continuous assessment. (pp108 – 124).* Uzopietro Publishers Company

Ekwonye, E.C.& Eguzo G.O. (2011). *Basic test theory in measurement and evaluation.* Joemankpa publishers.

Ercikan, K. & Koh, K. (2015). Construct comparability of the English and French versions of TIMSS. *International journal of testing (5), 23-35.*

Ernest, P, (2015). *International journal of education in Mathematics, science and technology.* 3(3), 187 – 192.

Henard, D.H, (2000). *Item response theory in reading and understanding more multivariate statistics, vol. II, Larry Grimm and Paul Yarnold, (Eds). American Psychological Association 67-97.*

Iweka, F. (2017). *Application of one parameter latent trait theory in the construct test items validity of Mathematics.Internation Journal of Interdisplinary Research Methods. 4(3)*, 11 – 22.

Jayanthi, E.C. (2014). *The application of an unfolding model of the PIRT type to measuremen of attitude. Applied psychological measurement. Vol. 12 pp.33 - 50*

Musselwhite, D.J. & Wesolowski, B.C. (2018). *standard error of measurement. In Bruce B. Frey (Ed) the SAGE encyclopedia of educational research, measurement and evaluation (pp 1588 – 1590)* Thousand Oaks: SAGE publication, Inc.

National Policy on Education, (2013). Federal Republic of Nigeria (2013), National Policy on Education 6th Edition.

Nworgu, B.G, (2015). *Educational measurement and evaluation theory and practice (2nd Ed).* University Trust Publisher.

Obinne, A.D.E. (2012). *Using IRT in determining test item prone to guessing.* Retrieved march 22, 2021, URI: http://dx.doi.org/wje.v2 nIp91.

Obinne, A.D.E. (2018). Test item validity: Item Response Theory (IRT) perspective for Nigeria. *Research journal in Organizational Psychology & Educational Studies* 2(1). Retrieved April, 30, 2021, from www.emergingresources.org

Otumegwu, T.U., Ezechukwu, I.R., Ojedapo, D.O., (2024). Development of multiple choice test in mathematics using item response theory. *Journal of Innovations in Educational Assessment, 6(1),* 10-24.

Reeve, B.B. (2000). *Item and scale-level analysis of clinical and non-clinical sample responses to the MMP1-2 depression scales employing Item Response Theory.* Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Ukozor, F.I. (2016). Validity of a test. In Ogomaka, P.M.C, Ekwonye, E.C, Ukozor F.I, and Onah F.E, (Ed) *Measurement and evaluation: A comprehensive text for students and teachers*. (pp 218 – 255) Flashpoint Publishers