International Journal on Integrated Education
https://journals.researchparks.org/index.php/IJIE

e-ISSN : 26203502
p-ISSN : 26153785

An Application of Item Response Theory to Basic Science Multiple-Choice Test Development and Validation

Mercy I. Ani, David O. Ekeh
Educational Foundations department, Alex Ekwueme Federal University, Ndufu Alike Ikwo, Ebonyi State, Nigeria
merciful.ssit@yahoo.com, revdavemaria@gmail.com

Jayson A. Dela Fuente
Northern Negros State College of Science and Technology, Philippines
delafuentejayson89@gmail.com

Abigail C. Obodo
Science Education department, Engu State University of Science and Technology
Enugu State, Nigeria
abigailujo@gmail.com

Abstract: The study made use of instrumentation research design while item response theory was applied, to develop and validate Basic Science multiple-choice tests. 600 junior secondary school II students consisted of a sample that was randomly selected from 20 government co-education secondary schools in Udi education zone of Enugu State, Nigeria. The study was guided by six research questions. A 40-test item of Basic Science multiple choice test was constructed by the researchers and used to collect data. Three experts subjected the instrument to content and face validation to ensure its validity. Two of them were from the departments of Science education and educational foundations, respectively. A reliability index of 0.85, was realized. Analysis of the data that were generated, was carried out, using the maximum likelihood estimation technique of BILOG-MG computer programming. It revealed that 40-test items with the appropriate indices consisted of the final instrument developed and was used to assess the ability of students in Basic Science. The result of the study confirmed the reliability of the items of the Basic Science Multiple choice questions based on the three-parameter (3pl) model. The findings again revealed that the multiple-choice Basic Science test items were difficult and that there was differential item functioning in Basic Science among male and female learners. Recommendations that were in line with the findings were made, such as that: teachers and examination bodies should adopt and encourage IRT in the development of test instruments used in measuring the ability of students in Basic Science and other subjects.

Keywords: Achievement Test, Basic Science, Differential Item Functioning (DIF), Item Response Theory (IRT), Multiple Choices.

Introduction

Basic Science in the Nigerian educational system was due to the re-alignment/ restructuring made in the curricula for both primary Science and junior secondary school integrated Science. Integrated Science, which is presently known as Basic Science, is the initial Science subject a student is taught at the junior secondary school level. Learners at the junior secondary school level, are required to learn Basic Science as a requisite for studying the core Science subjects (Physics, Chemistry, and Biology) at the senior secondary school level (Obodo, Ani, & Neboh, 2021). To be successful in studying single Science subjects at the senior secondary school level, a student was required to be proficient in Basic Science at the junior secondary school level. Hence, to effectively monitor students' progress in Basic Science, there is a need for improvement in student's learning and assessment (Dela Fuente, 2019).

The relationship between learning and effective assessment can be maintained by observing the progress of students and passing that information back to them (Ani, 2015). Learning according to Dela Fuente (2021) and Harlen and Deakin-Crick, (2002) cannot be predicted and the necessity to make adaptive adjustments to instruction and impact on the learner's willingness, capacity to learn and desire through assessment is imperative. Assessment systematically collects reviews and uses data collected about educational programs for the improvement of student learning (Dela Fuente & Biñas, 2020; Ani, 2019). To enable students develop a profound understanding, knowledge, and the ability to apply it as a result of their educational experiences, assessment as a process, gathers and discusses data collected from multiple and diverse sources (Huba & Freed, 2000). Basic Science in the Nigerian educational system was due to the re-alignment/restructuring made in the curricula for both primary and integrated Science in junior secondary schools. Integrated Science, which is presently known as Basic Science, is the initial Science subject a student is taught at the junior secondary school level. Junior secondary school Students are required to study Basic Science as a requisite for studying Physics, Chemistry, and Biology as core Science subjects at the senior secondary school level (Obodo, Ani, & Neboh, 2021). The proficiency of a student in Basic Science at the junior secondary school level is the requirement to successfully study single Science subjects at the senior secondary school level. Hence, to effectively monitor students' progress in Basic Science, there is a need for improvement in students' learning and assessment. This idea of assessment that concerns Continuous Assessment is expected to be carried out at all levels of the educational system for all categories of learners could be seen in the National Policy on Education (NPE, 2013). Achievement tests can be used for this type of assessment.

By analyzing an individual's present performance, an achievement test is an examination designed to evaluate a person's knowledge in a specified area or diverse areas and measures what a learner has learned and how he has learned over some time (Dela Fuente, 2021; Malcolm,2003). It also, evaluates a learner's achievement and monitors students' learning, and also provides immediate feedback to both learners and their teachers. High levels of mastery and higher achievement scores indicate students' readiness for more advanced levels of instruction while low achievement scores show that there are subject areas the learner should improve on or that certain subjects should be repeated. Assessment instruments, like essay and objective tests, are utilized by the teacher according to the aims of the measurement. This study is primarily on objective tests, although other assessment instruments, like essay and objective tests, are usually administered by teachers, based on the aims and objectives of the measurement.

Learners' academic achievement in a given instruction is measured by assessment instruments like objective tests (Berondo & Dela Fuente, 2021). They include multiple choices, true/false statements, missing words, incomplete sentences. They are called objective tests because the items that compose them must have precisely predetermined correct responses, no matter what educational objective it assesses. In objective tests, Respondents in this kind of test are required to choose the best possible answer(s) from the choices provided (Okoro, 2006). A multiple-choice test item that presents examinee with a question comprising of about four possible answers, out of which he selects one. Multiple choice items consist of a stem, which is the beginning part of the problem that presents the problem to be solved; as well as a set of options which are the possible answers the person taking the test can choose from, the correct answer is called the key while the incorrect ones are referred to as the distracters. The stem also contains information that is relevant to the test which also includes an incomplete statement that is to be completed and a question the examinee is required to answer. The multiple-choice tests, which are mostly self-administered are generally much more objective and scorers can apply a scoring key that allows them to agree perfectly (Meredith, Joyce & Walter 2007). Multiple choice questions are made up of a stem and a set of options. The stem is the part that starts the item with the following characteristics: it shows the 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 possible answers that the examinee can choose from are called the options, while the correct answer is referred to as the key, and the incorrect answers are called distracters (Ebuoh, 2018). The competence of the students is assessed with the test scores derived from the multiple-choice questions. Since multiple-choice questions contain a scoring key that allows the person scoring it to agree perfectly, they are often used for self-assessment by applying a scoring key.

Criteria, like reliability, validity, objectivity, and usability must however be satisfied by assessment instruments (Anene & Ndubisi, 2003).

The reliability of an instrument is the degree to which the instrument is consistent with whatever it measures (Dela Fuente, 2021; Nworgu, 2015). This means that if a test like Basic Science were to be administered as many times as possible, it would be expected that responses generated would vary slightly after each trial due to measurement error. This implies that, for any measuring instrument, the reliability is inversely proportional to the degree of error and vice versa. How much an instrument measures that it is constructed to measure is referred to as validity (Nworgu, 2015). If a test measures correctly, the particular attributes it is supposed to measure, its validity is said to be high. The tendency of individuals who administer and score a test not to distort its scores due to bias is known as objectivity. A test's usability is measured by the ease with which it provides the teacher with definite instructions that can be put to use with little chaos or difficulty. Beyond the calculation of the above-mentioned four qualities of items of Basic Science measuring instruments, determination of the quality of the instrument require other indices like item difficulty, item discrimination, and distractors. However, the measurement theory used determines the methods for calculating the qualities of items of the instrument to be used. Classical Test Theory (CTT) and Item Response Theory (IRT) are the two main measurement theories used for the study.

Classical test theory (CTT), which is based on the True Score theory regards the observed score (X) as consisting of the true scores (T) and an error component (E) (Aniugwu, 2017). The observed score of an examinee is regarded as an estimate of the true scores of the examinee added or deducted from a measurement error that cannot be observed (Algina & Crocker et al. 2008). CTT is comparatively simple, easy to interpret, and also empirically, considers a group of examinees' ability to succeed on a specified item, However, CTT has some disadvantages, because the item difficulty undergoes variations, that depend on the sample of the examinees. Consequently, the results of examinees obtained between different tests are difficult to compare and it is difficult to determine the proportion of examinees that get an item correctly in a sample. From a sample whose mean the ability of the result changes from a sample that is high to a low one (Npkone, et al. 2001). The estimates of achievement tests in secondary schools can be described, using CCT, regardless of its limitation, which is usually corrected by. Item Response Theory (IRT) is also called the Modern Theory (Troy- Gerard, 2004).

Item Response Theory (IRT), is based on an examinee's likelihood of success on a silent variable. It assesses a student's performance by using item distributions. An educational measurement scale that reports students' ability on both item and total instrument levels can be developed by this theory. It also contains a ratio scale and samples independent attributes. It is a technique that describes and breaks down the relationship between the test performance of a person who takes a test and the silent trait behind the performance, into its parts (Henard, et al. 2000). This is done by looking at the student's performance and then distributing the items according to the test taker's likelihood of success on a silent variable. The statistics of the parameters are estimated and interpreted but not varied in different populations of persons while the parameters of the persons are not varied across items. The IRT model assumes that one or more abilities of an examinee can be used to predict or explain his/her performance.

The likelihood of obtaining correct answers is modeled by IRT with the use of three logistic functions; namely: the one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL). The one-parameter logistic model, allows each question to have an independent difficulty variable. In this way, the probability is obtained. The level of discrimination of each item is modeled by the two-parameter logistic (2PL) model between students with high and low abilities. There is however a third item parameter which is known as the pseudo-guessing parameter. This item parameter is added to the three-parameter logistic (3PL) model and it echoes the probability that an examinee with a very low trait level would answer an item correctly, by merely guessing. The implication, therefore, is that by merely guessing, a student can answer a question in an achievement test correctly, thereby, providing an answer to a fact about something without certainty (Obinne, 2012). The probability that an examinee will be able to provide a correct answer to a question with a difficulty index, discrimination index, and a guessing index is given by a Guessing parameter model This model assumes that the probability of a correct answer to a question can be estimated and that there is a relationship between the ability of an individual, which is linked to the three parameters (difficulty, discrimination, and guessing). Within the latent trait test model, A test’s internal validity is estimated following the statistical fit of each item within its latent trait. The uniformity and high magnitude of the item discrimination equally signify the fit to the model. The absence of errors in item scoring and uniformity and high magnitude of item discrimination are indicators of fit to the model which is implied and is also an indicator that the effect of guessing on the test scores is negligible. IRT models are widely applied in Assessment instruments such as Basic Science achievement tests which have found IRT models very helpful for understanding learners’ abilities through assessing their test performance. The instrument for assessment must be fair before any Basic Science achievement test is considered unbiased for all examinees. When two groups of equal ability obtain the same score on each question of a test concerning the trait measured by the test, the test instrument is said to be unbiased.

There have often been indications that items could function differently for different learners’ group when the results of different subgroups are compared statistical methods which indicate that items could function differently for different groups of students are referred to as Differential item functioning (Madu, 2012). If an item’s Item Response Function (IRF) varies for two groups in a Basic Science achievement test, the implication is that differential item functioning has occurred. This means that individuals from different subgroups (e.g., females and males) but of equal ability, do not have the same chance of getting the same score (Meredith, Joyce, and Walter, 2007). . Therefore, any instrument developed for measuring achievement tests in Basic Science that is deficient of the Basic qualities that a test instrument should have, may suffer from differential item functioning. One of the limitations of the CCT model, which emphasizes aggregate-level performance in analyzing Basic Science achievement tests, is that an item may be labeled as biased when no bias exists. This could occur even where a large p-value difference and item-by-group interaction exists. Based on the foregoing, the aim of this study is, however, predicated on ensuring objectivity by using the type of measurement theory that emphasizes item-level performance rather than aggregate-level performance.

Most Basic Science teachers based on observation, write down their test items or pick up published past questions, without considering the psychometric properties of the test. Assessment examinees are expected to be treated equally in every assessment, which is not the case with assessment instruments developed by teachers through Classical test theory, as such instruments are group dependent, item statistics such as item discrimination and item difficulty. The researchers to avoid the limitations of instruments developed under Classical test theory, designed this study using a modern measurement theory (IRT). This ensured objectivity while measuring learners’ scores while analyzing items of Basic Science Multiple choice tests. The question posed is, therefore: would the instrument development and validation of multiple-choice tests in Basic Science be influenced by Item response theory? To avoid the limitations of instruments developed under Classical test theory. The objective of the study is therefore to apply Item response theory in the development and validation of Basic Science multiple-choice tests for Junior Secondary School (JSS II) students in Udi education zone, of Enugu State, Nigeria, is.

Research Questions

The researchers posed the six research questions below, to guide this study.

1. What are the standard errors of measurement of the test items of the Basic Science multiple-choice test?
2. How do the items of the Basic Science multiple choice test fit the three-parameter logistic (3PL) model?
3. What are the difficulty parameters of the test items of the multiple choice test in Basic Science?
4. What are the discrimination parameters of the test items of multiple-choice tests in Basic Science?
5. What are the guessing parameters of the test items of the multiple choice test in Basic Science examinations?
6. What are the Differential item functioning of the test items of the multiple choice test in Basic Science with respect to gender?

Methodology

The study employed an instrumentation research design since it is geared towards developing an instrument in Basic Science. Instrumentation research design is applied when the main purpose of the study is solely, to develop and standardize an instrument whose different psychometric properties (validity, reliability, usability e.t.c) have been empirical, and determined (Ali, 2006). The design is suitable for this study, since, the researchers developed Basic Science Multiple choice test questions that were psychometrically analyzed. The population for the study consists of all the Junior Secondary School II students in public secondary schools in Udi Education Zone of Enugu State, Nigeria. A sample of 600 Junior Secondary School II students was randomly selected from 20 government co-education secondary schools in the zone. Six research questions guided the study. The 40-test item of the Basic Science Multiple choice test was developed by the researchers following the guideline outlined by Herman and Lynn (1985) which include: the identification of objectives to be used, creating the test description, developing a test blueprint, construction and review of an initial item, trial testing, field testing, determination of the statistical properties of the items, conducting reliability and validity studies for the final version of the test and preparation of administration guidelines for the test.

The items were constructed using a test blueprint developed using the Basic Science curriculum and Bloom’s (1956) educational objectives taxonomy. Kendall Coefficient of Concordance was used by two of the experts from the department of Science education and educational foundations, of Enugu State University of Science and Technology (ESUT), respectively to establish the content validity. The estimate of the reliability of 0.85 of the Basic Science multiple tests was determined through the Kuder-Richardson formula 20 (K-R 20). The final test was administered to the sampled students and scored right (1) or wrong (0). The K-R 20 helped to establish the internal consistency of the Basic Science Multiple Test. The research questions were answered using the maximum likelihood estimation technique of the BILOG-MG V3 of 3PL MODEL computer programming. The three-parameter model of item response theory was used to establish the magnitude of item difficulty and item discrimination parameters of the Basic Science Multiple Test.

Results

Research Question One: What are the standard errors of measurement of the test items of the Basic Science multiple choice test?

Table 1: Standard errors of measurement of the test items of the Basic Science multiple choice tests based on three-parameter logistic (3PL) model

Item	S.E	Item	S.E	Item	S.E	Item	S.E
1	0.22	11	0.09	21	0.09	31	0.10
2	0.14	12	0.13	22	0.15	32	0.09
3	0.05	13	0.07	23	0.14	33	0.06
4	0.11	14	0.08	24	0.09	34	0.16
5	0.09	15	0.15	25	0.06	35	0.10
6	0.36	16	0.33	26	0.16	36	0.20
7	0.09	17	0.08	27	0.24	37	0.07
8	0.10	18	0.07	28	0.07	38	0.12
9	0.08	19	0.08	29	0.05	39	0.27
10	0.08	20	0.06	30	0.58	40	0.12

Table 1 shows the standard errors of measurement of the test items of the Basic Science multiple choice questions based on the three-parameter logistic (3PL) model. Based on the data in table 1, all the items except item 30 have a standard error of 0.05 to 0.44. Therefore, thirty-nine (39) items (98%) had a standard error below 0.50 and one (1) item (2%) had a standard error above 0.50. A standard error below 0.50 indicates high reliability while a standard error above 0.50 indicates low reliability. This high reliability indicated consistency in measuring the student's ability in Basic Science.

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

Table 2: Fits statistics of Basic Science multiple choice test based on three parameter logistic (3PL) model.

Item	Chi.sq	Prob	Item	Chi.sq	Prob	Item	Chi.sq	Prob	Item	Chi.sq	Prob
1	79.4	0.00*	11	42.2	0.18	21	45.4	0.00*	31	51.9	0.20
2	57.1	0.16	12	77,4	0.00*	22	41.0	0.00*	32	1.38	0.13
3	31.5	0.03*	13	13.7	0.06	23	29.3	0.00*	33	52.1	0.00*
4	18.2	0.00*	14	40.0	0.00*	24	92.6	0.24	34	96.6	0.15
5	55.0	0.08	15	84.2	0.13	25	52.1	0.00*	35	46.7	0.05
6	35.2	0.00*	16	18.0	0.02*	26	45.5	0.02*	36	94.3	0.09
7	90.9	0.14	17	79.0	0.06	27	103.4	0.08	37	70.4	0.00*
8	76.0	0.00*	18	46.0	0.07	28	33.5	0.00*	38	67.3	0.12
9	43.9	0.03*	19	43.7	0.00*	29	26.1	0.09	39	37.6	0.00*
10	31.7	0.09	20	21.3	0.00*	30	31.4	0.07	40	179.9	0.00*

** Significant*

Table 2 shows the chi-square goodness-of-fit analysis for the items of the Basic Science multiple choice test based on the three-parameter logistic (3pl) model. The summary of the results revealed that the chi-square value linked with the probability value ranged from 0.00 to 0.24. Based on the data in Table 2, Twenty (21) items (53%) that is items 1, 3, 4, 6, 8, 9,12, 14, 16, 19, 20, 21, 22, 23, 25, 26, 28,33, 37, 39, and 40 did not fit the three-parameter model because the items were below .05 level of significance. Nineteen (19) items (47%) that is, items 2, 5, 7, 10, 11, 13, 15, 17, 18, 24, 27, 29, 30, 31, 32, 34, 35, 36, and 38 fitted the three-parameter model

e-ISSN : 26203502
p-ISSN : 26153785

because the items were above .05 level of significance. These items are not marked with an asterisk. This implies that 21 items were statistically significant while 19 items were not statistically significant. The criterion for all the items fit/misfit was determined at a .05 level of significance.

Research Question Three: What are the difficulty parameters of the items of the Basic Science multiple choice test?

Table 3: Item threshold values (difficulty estimates) of the items of the Basic Science multiple choice test based on three parameter logistic (3PL) model.

Item	Threshold	Item	Threshold	Item	Threshold	Item	Threshold
1	-0.27	11	0.13	21	-0.22	31	-0.35
2	-0.94	12	0.26	22	-1.12	32	-0.59
3	0.17	13	0.19	23	0.06	33	0.44
4	-0.71	14	-0.30	24	0,10	34	-0.16
5	-0.48	15	-0.60	25	0.44	35	-0.61
6	-2.10	16	-1.49	26	-0.64	36	-0.38
7	0.73	17	-0.59	27	-1.38	37	0.34
8	-0.52	18	0.04	28	0.08	38	-0.35
9	0.35	19	-0.73	29	-0.18	39	-1.15
10	-0.59	20	-0.14	30	-2.38	40	0.27

Table 3 shows that twenty-six (26) items (65%) are items 1, 2, 4, 5, 6, 8, 10 14, 15, 16, 17, 19, 20, 21, 22, 26, 27, 29, 30, 31, 32, 34, 35, 36, 38 and 39 within the b-value range of -3 to +3 had negative difficult estimates while seventeen (14) items (35%) that are items 3, 7, 9, 11, 12, 13, 18, 23, 24, 25, 28, 33, 37 and 40 within the b-value range of -3 to +3 had positive difficulty estimates. The negative estimates imply that 26 items are easy while 14 items are difficult. Based on this information, none of the items were rejected in terms of difficulty levels.

Research Question Four: What are the discrimination parameters of the test items of the Basic Science multiple choice tests?

Table 4: Item parameters of the test items of the Basic Science multiple choice tests based on three parameter logistic (3PL) model.

Item	Slope	Item	Slope	Item	Slope	Item	Slope
1	0.43	11	0.13	21	1.14	31	0.45
2	0.99	12	1.10	22	0.38	32	0.57
3	0.96	13	1.21	23	1.02	33	3.30
4	0.47	14	0.60	24	0.45	34	0.58
5	0.51	15	0.32	25	3.30	35	0.52
6	0.23	16	0.19	26	0.32	36	0.21
7	0.97	17	0.66	27	0.25	37	1.71
8	0.51	18	1.29	28	0.67	38	0.39
9	0.50	19	0.77	29	0.97	39	0.20
10	0.67	20	0.97	30	0.13	40	1.29

Table 4 reveals that eight (8) items (20%), that is items 6, 11, 15, 16, 26, 27, 36 and 39 within the value range of .01 - .34 indicated very low discriminating values, while eighteen (12) items (30%) that is items 1, 4, 5, 8, 9, 14, 22, 24, 31, 32, 34, and 35 within the value range of .35 - .64 indicated low discriminating values. Also, twenty (17) items (43%) that is item 2, 3, 7, 10, 12, 13, 17, 18, 19, 20, 21, 23, 28, 29, 30, 38 and 40 within the value range of .65 - 1.34 indicated moderate discriminating values and (37, 33 and 25) items (7%) had values of 1.71, 3.30 and 3.30 respectively, meaning that the three items had a very high discriminating attributes.

Research Question Five: What are the guessing parameters of the test items of the Basic Science multiple choice test?

Table 5: Guessing parameters of the test items of the Basic Science multiple choice test based on three parameter logistic (3PL) model.

Item	Asymptote	Item	Asymptote	Item	Asymptote	Item	Asymptote
1	0.02	11	0.07	21	0.16	31	0.10
2	0.00	12	0.32	22	0.00	32	0.07
3	0.05	13	0.00	23	0.24	33	0.00
4	0.01	14	0.04	24	0.10	34	0.01
5	0.00	15	0.00	25	0.00	35	0.00
6	0.02	16	0.00	26	0.15	36	0.00
7	0.18	17	0.17	27	0.05	37	0.25
8	0.00	18	0.09	28	0.00	38	0.09
9	0.02	19	0.00	29	0.00	39	0.03
10	0.00	20	0.00	30	0.13	40	0.30

Table 5 shows the guessing (asymptote) values of the items of Basic Science multiple-choice questions based on the three-parameter logistic (3pl) model. The data reveals that items ranged from 0.00 to 0.32. Based on the data in table 5, thirty-seven (37) items (93%) that is items fall within the c-value range of 0.00 to 0.20 which shows that the items were desirable and the probability of getting an answer correctly by mere guessing is low. Only three (3) items (7%) fall within the c-value range of 0.20 to 0.30 that is items 12, 23, and 40 which shows that the items were not very good and the probability of getting an answer correctly by mere guessing is high.

Research Question Six: What are the differential item functioning of the test items of the Basic Science multiple choice test with respect to gender.

Table 6: Model for group differential item functioning of the test items of the Basic Science multiple choice test.

Item	P		Chi.Sq		Ite m	P		Chi.Sq	
	M	F	M	F		M	F	M	F
1	0.76	6.0	5.0*	6.4*	21	0.32	0.00	9.0*	21.0*
2	0.58	0.01	6.6*	19.8*	22	0.00	0.46	99.8*	83.8*
3	0.00	0.00	40.0*	105.6*	23	0.25	0.83	10.1*	4.2*
4	0.72	0.70	10.7	10.7	24	0.99	0.00	0.9*	68.8*
5	0.61	0.00	6.3*	22.3*	25	0.23	0,30	10.4*	9.4*
6	0.00	0.00	49.4*	109.2*	26	0.00	0.00	141.8*	228.0*
7	0.06	0.59	6.2*	6.5*	27	0.32	0.00	9.2*	61.0*
8	0.00	0.00	30.1*	30.0*	28	0.02	0.02	18.1*	18.0*
9	0.10	0.00	13.2*	20.3*	29	0.00	0.00	99.8*	83.8*
10	0.29	0.01	9.5	9.5	30	0.00	0.00	92.8.*	242.6*
11	0.00	0.00	27.5*	147.4*	31	0.89	0.03	3.5*	16.8*
12	0.97	0.83	2.1*	4.2*	32	0.79	0.98	4.7*	2.0*
13	0.00	0.00	80.2*	134.3*	33	0.23	0.30	10.4*	9.4*
14	0.00	0.81	20.7	4.5*	34	0.00	0.00	26.3*	36.9*
15	0.04	0.00	16.1*	71.9*	35	0.9	0.92	3.0*	3.2*
16	0.00	0.00	107.2	107.2	36	0.00	0.00	141.6*	228.0*
17	0.85	0.57	4.0*	6.7*	37	0.00	0.19	22.8*	11.2*
18	0.55	0.00	15.2*	200.0*	38	0.72	0.00	5.3*	22.2*
19	0.00	0.00	23.8*	45.0*	39	0.00	0.00	68.2*	113.5*
20	0.24	0.00	10.4*	101.8*	40	0.24	0.07	10.4*	14.5*

Table 6 shows the adjusted threshold values for group differential item functioning of the test items of the Basic Science multiple choice test. From the data, the result indicated that Differential Item Functioning (DIF) effects were observed among 37 items (93%), indicating that the 37 items were identified as significantly exhibiting differential functioning among male and female students. Only three (3) items (7%) that are items 4, 10, and 16 were identified as not exhibiting differential functioning among male and female students. This refers to uni-dimensionality ability. It reveals that the item discriminations are uniform and substantial. The chi-square values were used to dictate the differential item effect.

Discussion of Findings

The results are in line with the current development that validity should be determined by a quantitative approach for more objectivity. The result in Table 1 is in agreement with Obinne (2008) that an S.E of 0.50 and below is described as high reliability, while an S.E above 0.50 is described as low reliability. This finding also agrees with Meredith *et al* (2007) that if the reliability coefficient increases, the standard error of measurement becomes smaller. This implies that the reliability of the instrument ensures the consistency of the test instrument. For any measuring instrument, the smaller the error, the greater the reliability while the greater the error, the smaller the reliability. The result of the study also indicates that 53% of the items did not fit the three-parameter model because the items were below the .05 level of significance. While nineteen 47% of the items fitted the three-parameter model because the items were above the .05 level of significance. The findings in Table 2 revealed that Twenty (21) items were statistically significant while nineteen (19) items were not statistically significant. This corroborates Adedoyin (2010) finding that used the chi-square test with a probability greater than an alpha level of 0.05 significant level to select items that fit the model.

From the findings of data collected for Research Question 3 on Table 3, twenty-six (26) items (65%) within the b-value range of -3 to +3 had negative difficult estimates while seventeen (14) items (35%) within the b-value range of -3 to +3 had positive difficulty estimates. The negative estimates imply that 26 items are easy while 14 items are difficult. Based on this information, none of the items were rejected in terms of difficulty levels. The finding agrees with (Chong, 2013) that the difficulty parameter or the threshold parameter value tells us how easy or how difficult an item is. The finding of this study corresponds with Obinne (2008) that negative difficulty estimates indicate that the items are easy while positive difficulty estimates indicate that the items are hard. The findings which revealed that the items were selected based on the b-value range of -3 to +3 correspond with (Baker, 2001) that theoretically, difficulty values can range from - 00 to + 00, in practice, difficulty values usually are in the range of - 3 to + 3. The result in Table 4 reveals that 20% of the items within the value range of .01 - .34 indicated very low discriminating values, while 30% within the value range of .35 - .64 indicated low discriminating values. Also, 43% of the items within the value range of .65 - 1.34 indicated moderate discriminating values and 7% of the items had values of 1.71, 3.30, and 3.30 respectively, meaning that the three items had a very high discriminating attribute. The discriminating parameter indicates how well an item discriminates between respondents below and above the item threshold parameter, as indicated by the slope of the item characteristics curves (Reeve & Fayers, 2005). This result is in agreement with the findings of Baker (2001) who described the range 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 findings of data collected for Research Question 5 in Table 5 reveal that items ranged from 0.00 to 0.32. This indicates that thirty-seven (37) items (93%) that are items fall within the c-value range of 0.00 to 0.20 which shows that the items were desirable and the probability of getting an answer correctly by mere guessing is low. Only three (3) items (7%) fall within the c-value range of 0.20 to 0.30 that is items 12, 23, and 40 which shows that the items were not very good and the probability of getting an answer correctly by mere guessing is high. This higher c-value range indicates that the probability of getting an answer by mere guessing is high. The finding, however, supported Kamiri (2010) observation that the lowest c-values, the better indicating a lower probability of getting the answer correct by mere guessing of low-ability examinees. Harris (2005) asserted 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. The finding in Table 6 shows that 37 items (93%), were identified as significantly exhibiting differential functioning among male and female students while three (3) items (7%) were identified as not exhibiting differential functioning among male and female students. This finding is supported by Davis (2002) who noted that sometimes items are found to behave differently in distinct groups such as gender or language (such as loading on different dimensions in a multi-dimensional factor analysis or having largely different mean item scores). In other words, two examinees with the same latent trait value but differing in other characteristics may have different probabilities of response. The findings were determined at a 0.05 level of significance.

Conclusions

Based on the result of the findings the following conclusions were drawn:

1. That thirty-nine (39) items indicated high reliability of the test items while one (1) item indicated low reliability.
2. That twenty-one (11) items fitted the three-parameter model while twenty-nine (29) items did not fit the three-parameter model.
3. That twenty-three (33) items indicated difficult items while seventeen (17) items indicated easy items.
4. That Eight (8) items indicated very low discriminating values, Ten (10) items indicated low discriminating values, twenty (20) items indicated discrimination moderate values and two (2) items indicated high discriminating values.
5. That thirty-five (35) items were considered desirable, meaning that the probability of getting an answer correctly by mere guessing is low while five were considered not very good, and the probability of getting an answer correctly by mere guessing is high.
6. The findings further revealed that items function differently in Basic Science among male and female students.

Recommendations

Based on the findings of the study the following recommendations were made:

1. The psychometricians and measurement experts should organize workshops to educate teachers on the implications of quality tests. They should as well train teachers to know about the modern measurement framework called IRT as well as the necessary interpretations involved.
2. The examination bodies and teachers should be encouraged to adopt (IRT) in developing test items used in measuring students' ability in Basic Science. Education ministries and universities should try and assist students who are interested to study research on item response theory to get the software and necessary computer packages.
3. It is imperative to determine how the items in an instrument fit the IRT parameter model, such as one-parameter, two-parameter, and three-parameter logistic models.
4. The differential item functioning effects of items should be properly determined in the test instrument to avoid gender differences.

References

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

2. Anene, G.U. & Ndubisi, OG (2003). Test development process. In B. G. Nworgu (Ed.), Educational measurement and evaluation: Theory and practice pp.110-122). Nsukka: University Trust Publishers.
3. Baker, F.B. (2001). The Basics of item response theory. (2nd ed.).United States of America: ERIC clearinghouse on assessment and evaluation.
4. Berondo, R.G. & Dela Fuente, J.A. (2021). Technology Exposure: Its Relationship to the Study Habits and Academic Performance of Students. *Utamax : Journal of Ultimate Research and Trends in Education, 3*(3), 125-141. https://doi.org/10.31849/utamax.v3i3.7280
5. Black, P.J. & William, D. (2009). Assessment and classroom learning. Assessment in education. 5, 7-74
6. Chong, H.Y. (2013). A Simple guide to the Item Response Theory (IRT) and Rasch modeling. Retrieved from March, 2013, from http:// www.creativewisdom.com.
7. Cherkesova, L.V. (2016). Problemy sovremennoi fundamentalnoi nauki [Problems of modern fundamental Science]. Moscow: Publishing house of the Academy of Natural History. [In Rus.]
8. Crocker, L. & Algina, J. (2008). Introduction to classical and modern test theory. Fort Worth: Harcourt Brace Jovanovich.
9. Davis, L.L. (2002). Strategies for controlling item exposure in computerized adaptive testing with polytomous scored items. Unpublished doctoral dissertation, of Texas at Autin.
10. Dela Fuente, J.A. (2021). Contributing factors to the performance of pre-service physical Science teachers in the Licensure Examination for Teachers (LET) in the Philippines. *Journal of Educational Research in Developing Areas, 2*(2), 141-152. https://doi.org/10.47434/JEREDA.2.2.2021.141
11. Dela Fuente, J.A. & Biñas, L.C. (2020). Teachers' competence in information and communications technology (ICT) as an educational tool in teaching: An empirical analysis for program intervention. *Journal of Research in Education, Science and Technology*, 5(2), 61-76.
12. Dela Fuente, J.A. (2019). Driving Forces of Students' Choice in specializing Science: a Science education context in the Philippines Perspective. *The Normal Lights*, 13(2), 225-250.
13. Dela Fuente, J.A. (2021). Facebook messenger as an educational platform to scaffold deaf students' conceptual understanding in environmental Science subject: A single group quasi-experimental study. *International Journal of Education*, 14(1), 19-29. doi:10.17509/ije.v14i1.31386
14. Dela Fuente, J.A. (2021). Implementing inclusive education in the Philippines. College teacher experiences with deaf students. *Issues in Educational Research*, 31(1), 94-110. http://www.iier.org.au/iier31/dela-fuente.pdf
15. Ebuoh, C.N. (2018). Effects of Analytical and Holistic Scoring Patterns on Scorer Reliability in Biology Essay Tests", World Journal of Education.
16. Federal Republic of Nigeria (FRN) (2013). National Policy on Education (4th ed.). Lagos: NERDC press.
17. Harlen, W. & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. In EPPI Centre (Ed.), Research evidence in education library (1.1 ed., pp. 153–).
18. Harris, D. (2005). Educational measurement issues and practice: comparison of 1-, 2-, and 3- parameter IRT models. DOI: 10.1111/j.1745-3992.1989.tb00313.x
19. Henard, D.H. (2000). Item response theory, in reading and understanding more -

e-ISSN : 26203502
p-ISSN : 26153785

multivariate statistics, Vol. II, Larry Grimm and Paul Yarnold, (Eds)., Washington, DC: American Psychological Association, 67-97.

20. Huba, M.E. & Freed, J.E. (2000). Learner-centered assessment on college campuses: Shifting the focus from teaching to learning. Boston, MA: Allyn & Bac
21. Karami, H. (2010). A Differential Item Functioning analysis of a language proficiency test: an investigation of background knowledge bias. Unpublished Master"s Thesis. University of Tehran, Iran.
22. Madu, B.C. (2012). Analysis of Gender-Related Differential Item Functioning in Mathematics Multiple Choice Items Administered by West African Examination Council (WAEC). Journal of Education and Practice. Retrieved
23. May, 15, 2012, from ISSN /2222.1735 (Paper) 2222-288X (Online) Vol 3(8).
24. Malcolm, T. (2003). An achievement test. Retrieved November, 20, 2013, from http://www.wisegeek.com/what-is-an- achievement-test.htm
25. Makama, G.A. (2013). Patriarchy and Gender Inequality in Nigeria: The Way Forward.European Scientific Journal June 2013 edition vol.9, No.17 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431
26. Meredith, D.G., Joyce, P.G., & Walter, R.B. (2007). Educational research: an introduction (8th ed.). United State of America: Pearson Press.
27. Nkpone, H.L. (2001). Application of latent trait models in the development and standardization of physics achievement test for senior secondary students. Unpublished doctoral dissertation, University of Nigeria, Nsukka.
28. Nworgu, B.G. (2015). Introduction to Educational Measurement and evaluation: theory and practice (2nd ed.). Nsukka: Hallman Publisher.
29. Obinne, A.D.E. (2008). Psychometric properties of senior certificate biology examinations conducted by West African Examinations council: Application of item response theory. Unpublished doctoral dissertation, University of Nigeria, Nsukka.
30. Obinne, A.D.E. (2012). Using IRT in determining test item prone to guessing. Reprieved June, 20, 2013, URL: http://dx.doi.org/wje.v2 n1p91.
31. Obinne, A.D.E. (2013). Test item validity: item response theory (IRT) perspective for Nigeria. Research Journal in Organizational Psychology & Educational Studies 2(1). Retrieved January, 28, 2014, from www.emergingresource.org
32. Obodo, A.C., Ani, M.I., & Neboh, P.O (2021). Effects of guided inquiry and lecture teaching methods on junior secondary school Basic Science students' academic achievement. Journal of scientific Research and Methods. Maiden Edition. 18-29.
33. Okoro, O.M. (2006). Measurement and evaluation in education. Uruowulu-Obosi: Pacific Publishers Ltd.
34. Onunkwo, G.I .N. (2002). Fundamentals of education measurement and evaluation. Owerri: Cape Publishers Int'l Ltd.
35. Troy-Gerard, C. (2004). An empirical comparison of item response theory and Classical test theory item/person statistics. Unpublished doctoral dissertation, University Texas A&M.
36. Reeve, B.B. (2002). An introduction to modern measurement theory. Bethesda, Maryland: National cancer institution.
37. Reeve, B.B. & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire items and scale properties. In P. Fayers and R.D.
38. Hays (Eds.), Assessing quantity of life in clinical trials: method of practice. (2nd ed.). USA: Oxford university press. Retrieved September, 11, from http://cancer. Unic.edu/research/faculty/display member-plone.asp? ID-694