INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Validation of Basic Science Achievement Test Using R Software

Dr. Ibiene T. Longjohn
Associate Professor, Department of Educational Psychology, Guidance and Counselling,
Ignatius Ajuru University of Education, Rumuolumeni, Rivers State, Nigeria
Njigwum, Allwell Sunny
Ph.D. Student, Department of Measurement and Evaluation
Ignatius Ajuru University of Education, Nigeria
Ibeh, Chioma Perpetua
Ph.D. Student,Department of Measurement and Evaluation,
Ignatius Ajuru University of Education, Nigeria

Abstract:
The study is aimed at validation of Basic Science Achievement test using innovative techniques (R software). The study employed an instrumentation and descriptive survey research design. The Basic Achievement Test (BAT) instrument was made up of four components, which was used for generation of 98 items. Survey method was used to collect data from 567 Junior secondary school three (JSS3) students for trial testing. Simple random sampling was used to select public secondary schools in Port Harcourt metropolis. Four research questions guided the study. The study data was pretested and was found to meet the assumption of normality and no outlier before item analyses was done. Item analyses on the 98-item multiple choice test were done based on CTT technique alone since IRT assumption of unidimensionality was violated. Items from CTT analysis with rpbs≥0.20 and 0.30≤p≤0.80 were considered good items and selected. The R software was employed for all analyses carried out in the study while DIMPACK 1.0 was used to confirm test of unidimensionality. The reliability of the test was established through Cronbach Alpha, Split-half and Kuder Richardson 20 statistics using R, which produced a reliability coefficient of 0.87 for both Cronbach alpha and KR20 and 0.89 for Slit-half reliability. The result revealed that 55 items were considered good while, 43 items were marked for elimination. The findings show that BAT is both valid and reliable and thus, is recommended for measuring students' proficiency in Basic Science in public secondary schools in Port Harcourt metropolis.

Keywords: *Validation, basic science, R software*

1. Introduction

A person cannot be truly educated without the knowledge of science embedded at some level of their journey to enlightenment. Education in its holistic nature interacts mutually with the world of science. Science is a discipline that studies about everything around us, no wonder it is linked with every human endeavour that can be studied either as a body of knowledge or as an approach to solving problems (Scientific method) (Njigwum&Agugoesi, 2019). Science is simply defined as the knowledge we gather about our environment. In elaborate terms, science is the intellectual and practical activity, a systematic study of the structure and behaviour of the physical and natural world (environment), through observation and experimentation (National Teachers' Institute-NTI, 2012).

Science at the junior secondary level, is studied as a unified discipline which draws content from the different but related science subjects, thus, called 'Integrated Science or Basic Science.' This interdisciplinary curriculum approach allows the learner to see the concepts and the methodological principles which unites the separate subject matters, thus, harmonizing the knowledge derived from the integration (Gbamanja, 1992 in Njigwum&Longjohn, 2019). They also affirmed that this approach gives the learner a firm foundation in science which enables him/her proceed to the separate science subjects such as Biology, Chemistry and Physics.

The job of a teacher is not complete until he has determined how much of his teaching contents and lesson objectives have been mastered and learnt as demonstrated in behavioural change by the students. To achieve this, tests are the tools to give feedback on what a person has learnt or an instrument to obtain feedback that will determine the presence or absence of a particular trait (Obowo-Adutchay, 2014). To this end, test is the assessment of an examinee's ability, performance or achievement in a given task or subject (Asuru&Longjohn, 2008 in Njigwum, 2019). There are three types of tests (Aptitude, Intelligence and Achievement) although, the type used commonly in the classroom is Achievement test. Achievement test are tests designed to measure the degree of attainment of educational objectives in a content or subject area (Orluwene, 2012). Thus, it a test that measures how much learning has taken place after a planned

programme of training or instruction. They are administered after students have undergone some training on a specified curriculum or syllabus (Awai &Njigwum, 2016).

Kaplan and Saccuzo (2009) opined that achievement test possesses the following unique characteristics:

- Evaluate the effects of a known or controlled set of experiences (i.e., content learnt).
- Evaluate the product of a course of training.
- Rely heavily on content validation procedures.

Achievement test occupies a central position in the education system as it generates data (information) to evaluate the entire teaching and learning process of the school. As such, it is crucial to construct a test of high quality- that is a test with high validity and reliability. The process of designing test with satisfactory psychometric properties (validity and reliability) based on test theories (classical test theories or item response theories) is called instrumentation research, which is the science of test construction (Kpolovie, 2010). To Okorodudu (2012), test construction or test development is a process of designing instruments or scales of measurements for determining and locating an individual's performance on a 'quantitative continuum' in a given subject matter, psychological and social attributes.

A test can be studied from different perspectives and the items in the test can be evaluated according to different theories. Two of such theories are the *Classical Test Theory* (CTT) and the *Item Response Theory* (IRT). These theories are the two major frameworks that are used in educational measurement to develop, evaluate, determine the reliability and validity of tests, as well as improve the quality of test items. These frameworks are based on different assumptions and use different statistical approaches. CTT was originally the leading framework for developing and analyzing standardized tests. Later, IRT was developed to compliment the role of CTT.

However, the study will be anchored mainly on the use of CTT techniques. CTT is based on the assumption that an examinee has an observed score and a true score. The observed score of a test-taker is usually seen as a combination of an estimate of the true scores of that test-taker, plus/minus some unobservable error. The true score reflects what the test-taker actually knows, but it is always contaminated by different sources of errors. CTT utilizes measures of item characteristics, item difficulty and item discrimination, the values of which are dependent upon the distribution of examinee proficiency within a sample. Although the assumptions upon which classical test theory is based allow it to be applied to an assortment of test construction situations, these same assumptions appear to create weaknesses in the classical test theory model. The CTT based statistical indices are easy to compute, manipulate and understand by lay persons, but they vary from sample to sample. CTT has its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton in Awopeju&Afolabi, 2016). While CTT has proven very useful in test development, the two statistics that form its cornerstones, item difficulty and item discrimination are both sample dependent.

2. Steps in Validation of an Achievement Test

Therefore, to validate a test the following steps are recommended by Asuru (2015), Orluwene (2012), Onunkwo (2005) and Bloom, Madaus and Hastings (1981). These steps include:

- Test planning
- State the objectives
- Outline content
- Develop a table of specification or test blue print.
- Determine format and length of test items
- Item writing
- Face and content validation
- Trial testing
- Item analysis
- Compute discriminative index
- Compute difficulty index
- Item selection, revision
- Estimating reliability of test

The steps above, was used to guide the development and validation of the 100-item multiple choice. A few of the steps will be highlighted below while others are embedded in the methodology and results of the study.

2.1. Test Planning

According to Asuru (2015), test planning includes all the preparatory processes in test construction. The processes include;

2.1.1. Stating the Objectives

Since the test is an achievement test measuring educational or behavioural objectives in the cognitive domain, bloom's taxonomies of educational objectives will be used. They include: *knowledge, comprehension, application, analysis, synthesis and evaluation.* However, because of the test format (multiple choice) and the level of students (Junior Secondary School Students) the test will measure more of the lower educational objectives such as; knowledge, comprehension and application (Asuru, 2015).

2.1.2. Outline Content

The content for this project is drawn from Basic Science syllabus from JSS1 to JSS 3. The content covers the four broad themes of Basic Science (National Teachers' Institute, 2012), they include;

- Learning About our environment
- Living and Non-living Things
- You and Energy
- Science and Development

2.1.3. Develop of a Table of Specification

The table of specification is a two-dimensional grid which consists of course content in one direction and the behavioural objectives or learning outcome (cognitive domain) in the other dimension (Ogunleye, 2000; Orluwene, 2012). The weighting of each cell can be determined subjectively (through personal judgment or objectively based on time spent (number of weeks or hours) teaching each topic (Asuru, 2015).

Below in table 1: Shows the table of specification for a 100-item multiple choice test in Basic Science for students in certificate class (JSS 3).

Content	**Behavioural Objective**			**Total**
	Knowledge	**Comprehension**	**Thinking**	
	68%	**28%**	**4%**	**100%**
Learning about our Environment	1, 2, 3, 26, 27,10, 19, 76, 96, 98, 15, 22, 82, 86, 87, 99, 100, 46.	4, 5, 75, 14, 20, 85, 95.	16	26 (26%)
Living and Non-Living Things	25, 36, 37, 49, 50, 66, 67, 97, 48,58, 78, 80, 6, 9, 11, 28, 31, 34, 38, 39, 41, 42, 45, 19, 76, 98, 7, 17, 18, 21, 71, 37.	24, 55, 43, 44, 77, 92, 93, 94.		37 (37%)
You and Energy	61, 62, 63, 68, 70, 13, 79, 64, 30, 32, 33, 47, 51, 56, 57, 65, 83, 88, 90.	12, 13, 59, 72, 74, 81, 35, 69, 23, 29, 52, 60, 84.	54, 73, 89.	36 (36%)
Science and Development	91.			1 (1%)
Total	68	28	4	100

Table 1: A Table of Specification for 100-Item Basic Science Achievement Test

2.2. Determine Format and Length of Test Items

The test developed is a 100 – item test and the format is a five-option multiple choice objective test type. The test was administered to 567 students in J.S.S.3 (UBE Section) and the time duration is 1hr 30 minutes (the standard time duration in Junior School Certificate Examination for 100 – item multiple choice objective test).

2.3. Item Writing

The actual couching or writing of the test items is based on the table of specification, mental level of the students, purpose of the test and characteristics of the testees etc.

2.3.1. Characteristics of the Testees

- School Type: Public
- Subject: Basic Science
- Class: JSS3
- No. of Testees: 567 students
- Average Age: 14-16 years
- Sex: Mixed
- Content: JSS curriculum

2.4. Item Analysis (Trial Testing)and Item Selection

Trial testing is a fault-finding process used to empirically determine the adequacy of each item. Item analysis is employed to revise and improve both items and the test as a whole (Ukwuije&Opara, 2012). Here, good items are selected based on the discriminative index and difficulty of the items. Mehrens and Lehmann in Onunkwo (2005) suggested that an item should be retained as long as it discriminates positively. According to Asuru (2015), items with discriminative index of zero and items with negative values are rejected. However, Obowu-Adutchay (2014) stated that items with difficulty range of 0.40-0.70 are considered ideal for selection. However, using the modern technique for item analysis, the CTT framework for item selection will be $rpbs \geq 0.20$ and $0.30 \leq p \leq 0.80$ as recommended by Metibemu (2016) and Mitee (2019).

2.5. Innovative Techniques in Test Validation Using R

R is a programming language and a free software environment for statistical computing and graphics that you can use to clean, analyze, and graph your data (Weston, & Yee, 2017). It is widely used by researchers from diverse disciplines to estimate and display results and by teachers of statistics and research methods. It's free, making it an attractive option, but does rely on programming code — instead of drop-down menus or buttons — to get the job done. Programming languages can be intimidating but, we use R for research and teaching, and we believe that the benefits far outweigh the time and effort needed to start.

One of the most powerful characteristics of R is that it is open-source, meaning anyone can access the underlying code used to run the program and add their own code for free. This means that R will always be able to perform the newest statistical analyses as soon as anyone thinks of them. Also, the R community is noted for its active contributions in terms of packages and has brought together a community of programming and stats nerds (a.k.a., useRs) that you can turn to for help. In addition, anyone can write their own R code, which means anyone can add to the huge list of R's tools. Programmers submit their code to R in the form of 'packages.' Some packages specialize in specific kinds of analyses, while other packages are much broader. For example, the 'psych' package by William R. Revelle can do anything from descriptive statistics to item-response theory to mediation analyses (Weston, & Yee, 2017). At the start of 2017, there are just under 10,000 packages available. And as soon as a new statistical approach is developed, someone will create a new package or add new tools to an existing package.

On the other hand, validation of test deals with determining the psychometric properties (Reliability and Validity) of items as well as the entire test. This is line with the view of Kpolovie (2010), who opined the science of test construction deals with the process of designing test with satisfactory psychometric properties (validity and reliability) based on test theories (classical test theories or item response theories). Validity which ensures a test measure what it purports to measure can be established for an achievement test in two ways; first, is *content validity* through table of specification and secondly, is *empirical validation* through item analysis (computing difficulty and discrimination indices of the test). Here, item analysis is empirically used to detect faulty items that need editing or outright rejection.

Therefore, the innovative approach of test validation which is advanced in this paper, will include how to construct a test with high validity and reliability using simple, cheap and easy to apply technique all in one software. As such, the rigour and cost of using multiple statistical software for test validation will be eliminated. For this sake of this paper, R software was employed to complete the basic operations for test validation using mainly Classical Test Theory (CTT); and some of the techniques were recommended by Oyeniran (2021). The following steps was employed:

2.5.1. Testing for Outliers

After preparing your data for analysis, it is important you check if your data is free from outliers- these are mistakes or extreme values that deviates from the dataset or observations. This extreme score could reduce the accuracy of the result which have a negative effect on the outcome of the study. The use of boxplot and qqpplot are two major ways to check for outliers. These can be computed in R using the 'car' package to produce the results below.

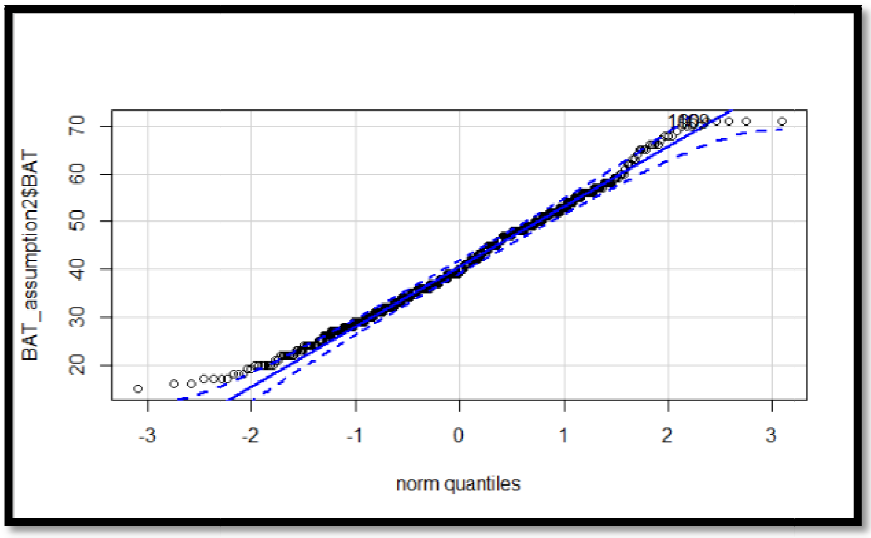

Figure 1: Qqplot of BAT Showing No Outliers and Normality

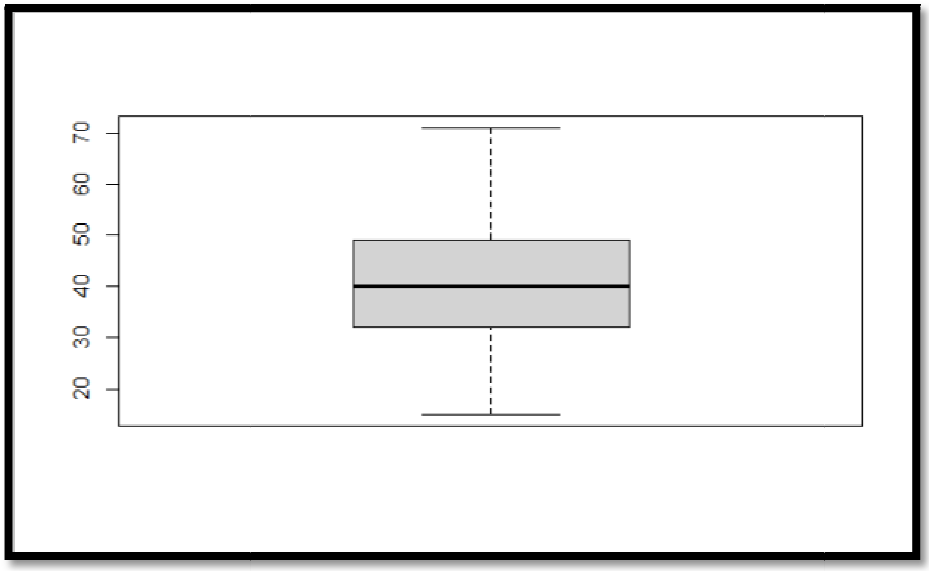

Figure2: Box Plot for BAT Showing No Outliers

2.5.2. Empirical Validation (Item Analysis)

This deals with computation of item difficulty and discrimination using either Classical test theory or Item Response Theory. This step is very important in test development, because a test is as good as the quality of the items that make up the test.

2.5.3. Reliability

Reliability is one psychometric indicator for the quality of an instrument or test. To compute reliability for a cognitive test, the following reliability statistics can be used: *KR20, Split-half and Cronbach Alpha.*

In conclusion, the ultimate goal of education is the training of the complete individual, who is deemed fit in character and in learning. This implies that a child who attends school is expected to show permanent changes in both cognitive and non-cognitive behaviours. Although, the primary aim of sending a child to school is to prepare him for a career which is tied to some subject content. However, it has been observed that most classroom teachers lack the expertise to produce valid tests (which help students achieve better) for their continuous assessment as well as end of term internal examinations. Thus, the major concern of this study is to help school teachers develop and validate good quality tests using easy, cheap and innovative technique. Therefore, the researchers deemed it necessary to carry out the task of validation of Basic Science Achievement test using innovative techniques (R software).

2.6. Research Questions

This study was guided by the following research questions:

- Does Basic Science Achievement Test (BAT) meet the IRT assumption of unidimensionality.
- What are the item parameters of the test (item calibration/estimation of item parameters) using CTT?
- How many of the items of the BAT survived under CTT framework?
- What is the reliability of the Basic Science Achievement Test BAT?

3. Methodology

The study used instrumentation and descriptive research design. Kpolovie (2010) defined instrumentation research as the science of test development- it is used for the purpose of test construction on the basis of test theories to ensure satisfactorily high validity and reliability as well as the most appropriate norm, criterion or domain in the measurement and evaluation of psychological attributes or human abilities. While descriptive research describes certain characteristics of the sample as they are at the time and it promotes collection of data from a large sample several characteristics (Nwankwo, 2016). The instrumentation design was used to develop and validate BAT instrument using Classical Test theory (CTT). However, descriptive survey design was used to collect data from a large sample public school students for pilot testing.

The target population for the study included all public Junior secondary school three students (JSS 3) in Port Harcourt metropolis in Rivers State. The population consists of about 5,315 students (Rivers State Ministry of Education, 2021) drawn from 14 schools in the region. The population is drawn from only public junior secondary school students.

Therefore, a sample of 567 students was used for the study. The sample size chosen exceeds the minimum sample size for the population estimated using the Krejcie and Morgan graph (*minimum sample of 360 for a population of 6000 persons*) cited in Kpolovie (2011). Simple random sampling technique was adopted to draw six secondary schools from the 14 public secondary schools within the target region. As such, intact classes of 100 students were randomly selected from each of the six public schools to make up the sample for the study.

The instrument for data collection is Basic Science Achievement Test (BAT). The BAT is a mix of self-developed Multiple-choice items and items drawn from past Basic Education Certificate Examination (BECE) in Rivers State. The test

is composed of 100 items, each item consists of a stem and five options (A, B, C, D and E). Correct response attracted a score of 1, while incorrect response attracted 0.

Data for this study was collected directly through the help of subject teachers who were used to administer the test as an internal test. Out of 600 tests instrument shared, 567 tests papers (i.e., 94.5%) were retrieved as completed test for the analysis.

The validity of the test which ensures a test measure what it purports to measure was established using two methods; first, is *content validity* through table of specification and secondly, is *empirical validation* through item analysis (computing difficulty and discrimination indices of the test). The content validity was established using table of specification. This is in agreement with Okorodudu (2012) and Kpolovie (2010), they affirmed that validity of any achievement test can best or most appropriately be ascertained or established through content validity using test blue print. Also, the face and content validities of the test was further established by subject specialists (using senior teachers in Basic Science). The experts were provided with the test form and test blue print for editing and vetting of the instrument. All suggestions and corrections made were effected before trial testing.

The reliability of the test was established through Cronbach Alpha, Split-half and Kuder Richardson 20 statistics using R, which produced a reliability coefficient of 0.87 for both Cronbach alpha and KR20 and 0.89 for Slit-half reliability. This coefficient is high enough to affirm the reliability of the test, since it is above the 0.7 benchmark for a reliable instrument.

For method of data analysis, SPSS and MS Excel software were employed for data coding, data cleaning and conversion to file csv. formats for R environment. DIMPACK 1.0 software was utilized for Unidimensionality analysis while, different packages in R were employed for item analysis using CTT techniques.

4. Results

4.1. Research Question One

Does Basic Science Achievement Test meet the assumption of unidimensionality. To test this, the Bootstrap Modified Parallel Analysis Test (BMPAT), implemented by the function unidimTest in the ltm r-package and the Stout's Test of Essential Unidimensionality (STEU) implemented in DIMTEST were used. First the BMPAT method was used and subsequently, the results were cross-validated with the STEU. The results are presented as follow:

4.1.1. UnidimensionalityTest with BMPAT

The results of the dimensionality test with the BMPAT are presented in Table 2.

	Values	**p-value**
Second eigenvalue in the observed data	4.4401	0.005
Average of second eigenvalues in Monte Carlo samples	2.6361	
Monte Carlo samples	200	

Table 2: Bootstrap Modified Parallel Analysis Test (BMPAT) of Unidimensionality

The results presented in Table 1 shows that the second eigenvalue of the observed data (4.4401) is larger than the second mean eigenvalues of the simulated data (2.6361). Also, the observed difference was statistically significant (p = 0.005). This implies that the test items are not unidimensional, suggesting that the null hypothesis should be rejected. The cross-validation of the result with the STEU in the next section will be used in the final determination of dimensionality.

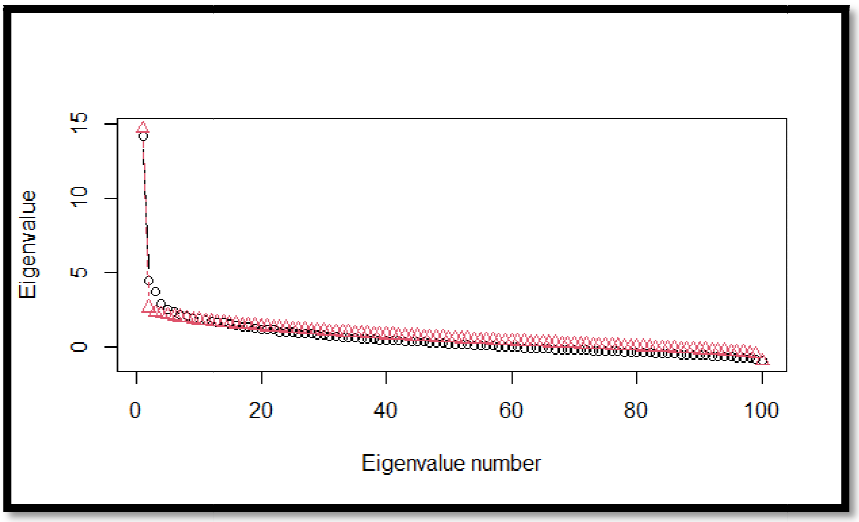

Figure 3: BMPAT Scree Plot

Figure 3 shows the scree plot of the BMPAT analysis. From the plot the eigenvalues of the second factors of both the test data and the mean eigenvalues of the simulated sample are displayed. The plot also shows that the second

eigenvalue in the observed data is substantially higher than the mean of the second eigenvalues of the simulated data. This provides graphical visualization for the result in Table 4, and further suggests that the test data did not fulfill the unidimensionality assumption.

4.1.2. Unidimensionality test with DIMTEST

To perform the Stout's test, the items were divided into two subtests that are as dimensionally distinct as possible, the Partitioning Subtest (PT) and the Assessment Subtest (AT) using DIMPACK 1.0. The null hypothesis being tested is that, the responses are unidimensional (the average covariance within groups = 0), this, non-rejection of the null Hypothesis indicates that the assumption of unidimensionality is tenable and vice versa. Table 3 presents the result

TL	TGbar	T	p-value
11.1250	7.8862	3.2227	0.0006

Table 3: Stout Test of Essential Unidimensionality (STEU) for BAT

Table 3 shows that the AT were dimensionally distinct from each of the remaining items of the test. The Stout statistic T = 7.8862 ($p < 0.05$) was significant, indicating that the average covariance within groups is not zero, hence the hypothesis of unidimensionality was rejected. This showed that there was more than one underlying trait that accounted for the variation observed in students' responses to the test items, the assumption of unidimensionality of the test items did not hold and the study does not favour the application of multidimensional method, therefore, only Classical Test technique will be employed for item calibration.

4.2. Research Question 2

What are the item parameters of the test (item calibration/estimation of item parameters) using CTT?

To answer this question, the test items were subjected to psychometric analysis using R programming language. Table 4 presents the CTT statistics, including ***p*** which represents the item difficulty indices and the item discrimination indices.

Sample	Difficulty	Remark	Discrimination	Remark
ITEM1	0.712032	GOOD	0.295858	GOOD
ITEM2	0.854043	BAD	0.242604	BAD
ITEM3	0.534517	GOOD	0.449704	GOOD
ITEM4	0.203156	GOOD	0.201183	GOOD
ITEM5	0.424063	GOOD	0.319527	GOOD
ITEM6	0.566075	GOOD	0.473373	GOOD
ITEM7	0.556213	GOOD	0.497041	GOOD
ITEM8	0.952663	BAD	0.094675	BAD
ITEM9	0.30572	GOOD	0.171598	BAD
ITEM10	0.47929	GOOD	0.39645	GOOD
ITEM11	0.512821	GOOD	0.35503	GOOD
ITEM12	0.323471	GOOD	0.159763	BAD
ITEM13	0.698225	GOOD	0.467456	GOOD
ITEM14	0.641026	GOOD	0.47929	GOOD
ITEM15	0.242604	BAD	0.195266	BAD
ITEM16	0.285996	BAD	0.106509	BAD
ITEM17	0.682446	GOOD	0.455621	GOOD
ITEM18	0.650888	GOOD	0.544379	GOOD
ITEM19	0.747535	GOOD	0.325444	GOOD
ITEM20	0.42998	GOOD	0.485207	GOOD
ITEM21	0.656805	GOOD	0.366864	GOOD
ITEM22	0.248521	BAD	0.12426	BAD
ITEM23	0.504931	GOOD	0.266272	GOOD
ITEM24	0.475345	GOOD	0.236686	GOOD
ITEM25	0.609467	GOOD	0.43787	GOOD
ITEM26	0.741617	GOOD	0.461538	GOOD
ITEM27	0.852071	BAD	0.272189	GOOD
ITEM28	0.518738	GOOD	0.390533	GOOD

Sample	Difficulty	Remark	Discrimination	Remark
ITEM29	0.313609	GOOD	0.313609	GOOD
ITEM30	0.100592	BAD	0.076923	BAD
ITEM31	0.771203	GOOD	0.319527	GOOD
ITEM32	0.453649	GOOD	0.52071	GOOD
ITEM33	0.209073	BAD	0.12426	BAD
ITEM34	0.370809	GOOD	0.331361	GOOD
ITEM35	0.485207	GOOD	0.532544	GOOD
ITEM36	0.191322	BAD	0.047337	BAD
ITEM37	0.309665	GOOD	0.218935	GOOD
ITEM38	0.25641	BAD	0.142012	BAD
ITEM39	0.658777	GOOD	0.491124	GOOD
ITEM40	0.282051	BAD	0.254438	BAD
ITEM41	0.412229	GOOD	0.491124	GOOD
ITEM42	0.656805	GOOD	0.467456	GOOD
ITEM43	0.648915	GOOD	0.295858	GOOD
ITEM45	0.751479	GOOD	0.39645	GOOD
ITEM46	0.157791	BAD	0.16568	BAD
ITEM47	0.067061	BAD	0.071006	BAD
ITEM49	0.327416	GOOD	0.390533	GOOD
ITEM50	0.197239	BAD	0.236686	BAD
ITEM51	0.209073	BAD	0.112426	BAD
ITEM52	0.153846	BAD	0.088757	BAD
ITEM53	0.096647	BAD	0	BAD
ITEM54	0.236686	BAD	0.213018	BAD
ITEM55	0.122288	BAD	0.053254	BAD
ITEM56	0.635108	GOOD	0.195266	GOOD
ITEM57	0.293886	BAD	0.201183	BAD
ITEM58	0.455621	GOOD	0.455621	GOOD
ITEM59	0.526627	GOOD	0.35503	GOOD
ITEM60	0.230769	BAD	0.230769	BAD
ITEM61	0.623274	GOOD	0.408284	GOOD
ITEM62	0.431953	GOOD	0.195266	GOOD
ITEM63	0.22288	BAD	0.230769	BAD
ITEM64	0.358974	GOOD	0.349112	GOOD
ITEM65	0.175542	BAD	0.118343	BAD
ITEM66	0.420118	GOOD	0.349112	GOOD
ITEM67	0.38856	GOOD	0.266272	GOOD
ITEM68	0.252465	BAD	0.130178	BAD
ITEM69	0.571992	GOOD	0.426036	GOOD
ITEM70	0.248521	BAD	0.130178	BAD
ITEM71	0.254438	BAD	0.159763	BAD
ITEM72	0.368836	GOOD	0.213018	GOOD
ITEM73	0.13215	BAD	0.071006	BAD
ITEM74	0.29783	BAD	0.171598	BAD
ITEM75	0.357002	GOOD	0.118343	BAD

Sample	Difficulty	Remark	Discrimination	Remark
ITEM76	0.579882	GOOD	0.360947	GOOD
ITEM77	0.317554	GOOD	0.189349	BAD
ITEM78	0.335306	GOOD	0.153846	BAD
ITEM79	0.234714	BAD	0.136095	BAD
ITEM80	0.485207	GOOD	0.372781	GOOD
ITEM81	0.353057	GOOD	0.04142	BAD
ITEM82	0.351085	GOOD	0.230769	GOOD
ITEM83	0.295858	BAD	0.059172	BAD
ITEM84	0.25641	BAD	0.035503	BAD
ITEM85	0.353057	GOOD	0.272189	GOOD
ITEM86	0.441815	GOOD	0.284024	GOOD
ITEM87	0.207101	BAD	0.337278	GOOD
ITEM88	0.238659	BAD	0.100592	BAD
ITEM89	0.143984	BAD	0.059172	BAD
ITEM90	0.38856	GOOD	0.325444	GOOD
ITEM91	0.404339	GOOD	0.307692	GOOD
ITEM92	0.189349	BAD	0.159763	BAD
ITEM93	0.205128	BAD	0.153846	BAD
ITEM94	0.315582	GOOD	0.236686	GOOD
ITEM95	0.431953	GOOD	0.461538	GOOD
ITEM96	0.623274	GOOD	0.455621	GOOD
ITEM97	0.473373	GOOD	0.378698	GOOD
ITEM98	0.268245	BAD	0.195266	BAD
ITEM99	0.439842	GOOD	0.177515	BAD
ITEM100	0.605523	GOOD	0.443787	GOOD

Table 4: Item Parameter Estimates and Survived Items under CTT

From Table 4, it was observed that the CTT framework gave the estimates of all the item parameters of the 98 items subjected to its item analysis process.

4.3. Research Question 3

How many of the items of the BAT survived under CTT framework?

To answer the research question, we examine Table 5. Looking at the columns for the item difficulty and discrimination, the items that survived the set criteria ([a] $0.20 \leq p \leq 0.80$ and [b] $r_{pbs} \geq 0.20$) are rated as being good (see Remarks column). The items to be deleted based on the CTT analysis are presented in Table 4.

Item parameter	Number Deleted	Items Deleted
Difficulty	**37**	Items 2, 8, **15**, 16, 22, **27,** 30, 33, 36, 38, 40, 46, 47, 50, 51, 52, 53, 54, 55, 57, 60, 63, 65, 68, 70, 71, 73, 74, 79, 83, 84, **87**, 88, 89, 92, 93, 98.
Discrimination	**38**	Items 2, 8**, 9, 12**, 15, 16, 22, 30, 33, 36, 38, 40, 46, 47, 50, 51, 52, 53, 54, 55, 57, 60, 63, 65, 68, 70, 71, 73, 74, 79, 83, 84, 88, 89, 92, 93, 98, 99.
Total from both	43	

Table5: Items Deleted under CTT Framework

From Table 5, based on the criteria for deletion of items, 43 items, were adjudged bad items and are marked for deleting from the pool of items. Here, good items are selected based on the discriminative index and difficulty of the items. However, using the modern technique for item analysis, the CTT framework for item selection was *rpbs≥0.20 and*

0.30≤p≤0.80 as recommended by Metibemu (2016) and Mitee (2019). Where, *p* stands for item difficulty and *rpbs* stands for item discrimination indices (point biserial correlation). Therefore, out of the 98 items developed for the test, 43 items were rejected while, 55 items were accepted to make the final test form.

4.4. Research Question 4

What is the reliability of the Basic Science Achievement Test BAT?

The reliability of the test was determined using three reliability methods, namely; Kuder-Richardson–20, Split-half method and Cronbach alpha method. All the reliability methods utilized single testing or administration; the scores generated were used to compute the reliability coefficient. After computation, the following coefficients were obtained as the measure of internal consistency of the test:

Reliability Method	Coefficients
KR_{20}	0.87
Split-Half	0.896
Cronbach Alpha	0.87

Table 6: Reliability Coefficients

This coefficient is high enough to affirm the reliability of the test, since it is above the 0.7 benchmark for a reliable instrument.

5. Discussion

To answer the first research question, the Bootstrap Modified Parallel Analysis Test (BMPAT), implemented by the function unidimTest in the ltm r-package and the Stout's Test of Essential Unidimensionality (STEU) implemented in DIMTEST were used. The result showed that both the BMPAT and the STEU proved that Basic Science Multiple Choice test violated the assumption of unidimensionality. This showed that there was more than one underlying trait that accounted for the variation observed in students' responses to the test items, the assumption of unidimensionality of the test items did not hold, hence the test is multidimensional. The result of the study is line with Alade, Aletan and Sokenu(2020), and Oye (2021) who agreed that their test items violated the IRT assumption of dimensionality. On the contrary, the work of Mitee (2019), and Ogbonna (2018) disagrees with the present study. This may be due to the length of items used in the present study as well as the broad nature of the content sampled (Basic science from JSS1 to JSS 3) implying that test items are homogenous and that only one latent trait is responsible for examinee performance. However, since the assumption of unidimensionality of the test items did not hold and that the application of multidimensional method is outside the scope of the present study, therefore, only Classical Test technique was employed for item calibration.

To answer the second and third objectives of the study, item calibration was computed under Classical Test Theory (CTT) framework and good items were selected based on the discriminative index and difficulty of the items. Also, out of the 98 items subjected for analysis, 43 items were rejected while, 55 items were accepted to make the final test form. From the results it was observed that the difficulty for the 55 selected items ranged from 0.20 (item 4) to 0.77 (item 31); while the item discrimination ranged between 0.200 (item 2) and 0.53 (item 42). Item discrimination indices on the other hand ranged between 0.064 (Item 2) to 0.317 (Item 42). The test items are seen to have moderate difficulty indices (about 71% of the selected items), which indicates that the items were very good under the CTT framework, and also the items are shown to be highly discriminating under CTT.

Finally, the reliability of the test was determined using three reliability techniques namely; Kuder-Richardson–20, Split-half method and Cronbach alpha method. These techniques were recommended for assessing the reliability of cognitive instrument (Onunkwo, 2002; Asuru, 2015; Bloom, Madaus& Hastings, 1981). After computation, the following coefficients were obtained as the measure of internal consistency of the test from three reliability methods, they are: KR20=0.87, Split-Half= 0.896 and Cronbach Alpha= 0.87. These coefficients are high enough to affirm the reliability of the test, since they rose above the 0.7 benchmark for a reliable instrument. More so, the close reliability coefficients of the test showed its stability across the several reliability techniques used.

6. Conclusion

The result revealed that 55 items were considered good items while, 43 items were marked for elimination. The reliability of the BAT ranged between 0.87 to 0.89. The findings show that the BAT is both valid and reliable and thus, is recommended for measuring students' academic achievement in Basic Science. Therefore, the BAT will be useful for teaching and learning, research purposes and evaluation of Basic Science education.

7. Recommendations

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

- BAT can be used as a valid test for measuring students' proficiency in Basic Science in public secondary schools in Port Harcourt metropolis.
- Basic science teachers can adopt BAT as MOCK test for Basic Education Certificate Examination (BECE).
- Teachers and researchers should be trained to explore R software to develop and validate their tests.
- Educators and experts should provide resources on the use of advanced techniques for validation of test such IRT approach using R.

8. References

i. Alade, O. M., Aletan, S. &Sokenu, B. S. (2020) investigating the extent to which the 2018 West Africa Senior Secondary Certificate Examination Mathematics objective tests. *African Journal of Behavioural and Scale Development Research,AJB-SDR*, 2(1), 8-16.

ii. Asuru, V. A. (2015). *Measurement and evaluation in education and psychology* (2nded.). Port Harcourt: Pearl Publishers International Ltd.

iii. Awai, D. &Njigwum, A. S. (2016). Predictive Validity of JSCE Integrated Science on SSCE Chemistry Performance in Public Secondary Schools in Rivers State. *Icheke Journal of the Faculty of Humanities,* 14 (2), 293-309.

iv. Awopeju. O. A., &Afolabi, E. R. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal*, 12(28), 263-270. http://dx.doi.org/10.19044/esj.2016.v12n28p263

v. Bloom, B. S., Madaus, G. F. & Hastings, J. T. (1981). *Evaluation to improve learning*. New York: McGraw Hill, Inc.

vi. Kaplan R. M. &Saccuzzo, D. P. (2009). *Psychological Testing Principles, Applications and Issues* (7th ed.). United States: Thomson Wadsworth.

vii. Kpolovie, P. J. (2010). *Advanced research methods*. Owerri:Springfield Publishers Ltd.

viii. Kpolovie, P. J. (2011). *Statistical Techniques for Advanced Research*. Owerri:Springfield Publishers Ltd.

ix. Metibemu, M.A. (2016). Comparison of classical test theory and item response theory in the development and scoring of senior secondary school physics tests in Ondo State. Unpublished Ph.D thesis. Institute of Education. University of Ibadan.

x. Mitee, T. L. (2019). Comparative study of classical test theory and item response theory using item analysis results of quantitative chemistry achievement test. *African Journal of Behavioural and Scale Development Research* (*AJB-SDR)*, 1(1), 26-36.

xi. National Teachers' Institute –NTI (2012). *Manual for retraining of JSS teachers in Basic Science.*Kadunna: NTI Press. pp.1-4.

xii. Njigwum, A. S. (2019). Predicting Senior School Certificate Examination Performance in Mathematics and English Language from the Junior School Certificate Examination Performance in Obio/Akpor Local Government Area. Unpublished MEd Thesis, Ignatius Ajuru University of Education.

xiii. Njigwum, A. S. &Agugoesi, O. J. (2019). Junior Secondary Certificate Integrated Science Performance as a Predictor of SSCE Physics Performance in Public Secondary Schools in Port Harcourt Metropolis. A paper presented at 21st Annual Conference of the Association of Educational Researchers and Evaluators of Nigeria (ASSEREN), 22nd – 26th July at ObafemiAwolowo University, Ife, Nigeria.

xiv. Njigwum, A. S. &Longjohn, I. T. (2019). Junior Secondary Certificate Basic Science Performance as Predictor of Senior School Certificate Examination (SSCE) Biology Performance in Public Secondary Schools in Port Harcourt Metropolis. *African Journal of Behavioural and Scale Development Research (AJB-SDR),* 1(1), 101-107.

xv. Nwankwo, O.C. (2016). *A Practical Guide to Research Writing* (Rev. 6th ed.). Port Harcourt: M & J Grand Orbit and Communication Ltd. pp.71-73.

xvi. Obowu-Adutchay, V. (2014). Test development. In Obagah, M.O.N &Inko-Tariah, D. C. (Eds). *Educational measurement and evaluation*. Port Harcourt: Rodi Printing and Publishing Company. pp. 87-107.

xvii. Ogbonna, J. U. (2018). Application of Three- Parameter Latent Trait Model in The Development and Validation of Mathematics Achievement Test. Unpublished Ph.D Thesis, University of Port Harcourt.

xviii. Ogunleye, A. O. (2000). *An introduction to research methods in education and social sciences*. Lagos: Sunshine International Publication (Nig.) Ltd.

xix. Okorodudu, R. I. (2012). *Understanding educational and psychological measurement and evaluation (with CTT, GTT and IRT Theories*). Abraka:University printing press, Delta State University.

xx. Onunkwo, G. I. N. (2005). *Continuous assessment for Nigerian scholars.* Onitsha:Vigo publishers International.

xxi. Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation*. Owerri:Cape publisher International Limited.

xxii. Orluwene, G. W. (2012). *Fundamentals of testing and non-testing tools in educational psychology.* Port Harcourt: Harey Publications Coy.

xxiii. Oye, P. N. G. (2020). Assessment of Item Parameters of Rivers State Basic Education Certificate Examination (BECE) Mathematics Objective Items Using IRT and CTT Techniques. Unpublished PhD Thesis, Ignatius Ajuru University of Education.

xxiv. Oyeniran, D. O. (2021). Adoption of Innovative Approaches for Cognitive Test Validation. AB-ReAP Workshop on 'Skills Development for Instrumentation and Statistical Tools Adoption for Ground-Breaking Research' held at Ignatius Ajuru University of Education, Port Harcourt, between 11- 14th May, 2021.

xxv. Ukwuije, R.P.I. &Opara, M. I. (2012). *Test and measurement for teachers* (3rd ed.). Port Harcourt: Chadik Printing Press. pp.129-130.

xxvi. Weston, S. J., & Yee, D. (2017). Why you should become a useR: A brief introduction to R. *APS Observer*, *30*(3).