ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Construction and Standardization of Economics Achievement Test Using Classical Testing and Item Response Methods

[1]Dr. (Mrs) O.M. Alade, [2]Dr. 'Sola Aletan, [3]Dr. A. Osoba
[1,2] Department of Educational Foundation, University of Lagos
[3] The West African Examinations Council, Lagos

Abstract
The paper reports construction and standardization of the Economics Achievement Test (EAT). Multi-stage sampling process involving simple random, stratified and purposive sampling techniques were used for this study. Five research questions guided the study. A self-constructed valid and reliable 120-item Economics Achievement Test (EAT) was administered on 1,200 students. The items were pilot tested on similar sample but were not part of the study. Students' sex, school location (urban or rural) and school type (publicly or privately owned) were reflected in the selection of participants to ensure representativeness. The instrument was trial tested and the Kuder-Richardson (K-20) reliability consistency was .799. The study made use of Iteman-4, range and BILOG to analyze data. Results show that the number of items generated was higher in IRT (77 items) than CTT (65 items). It is recommended that in the construction, standardization and selection process, both CTT and IRT approaches should be employed, and the standardized Economic Achievement Test will be a good instrument in preparing students for either internal or external assessments and in comparing their performance.

Keywords: *Achievement, Economics, CTT, IRT, Standardized Testing*

Introduction

Tests are used to examine students' knowledge of the subject matter in a bid to determine what has been learned and to measure the levels of skill and knowledge that have been reached. Testing is a procedure for observing individuals and describing them numerically or categorically. Testing is also a method of measuring behaviour of individuals by presenting a set of questions in an orderly manner. There are different kinds of tests used for various purposes in educational settings. Achievement tests are intended to scale cognitive abilities, understanding and talents acquired, as well as academic progression. Aptitude tests are more specific and used to appraise ability in a specific field to forecast future performance. Above all, any test to be used must be standardized to produce valid and reliable results.

Tests have always been a great companion for teachers. Teachers use different tests in their classroom routines. Students are tested regularly, sometimes, monthly, termly and even annually. In the classroom, tests are used for measuring students' cognitive ability and knowledge in a particular subject. Kimau (2018), views the purpose of any tests among others to include: ascertaining what students have learned; identifying students' strengths and weaknesses; providing platforms for awards and recognition; providing ways to measure teachers' and schools' effectiveness.

Tests can either be teacher-made or standardized. When a test is developed and designed by the teachers is referred to as teacher-made while standardized tests are carefully constructed by public examining bodies or commercial outlets such that there is consistency of procedure in the administration, marking and analysis of the test results. Standardized tests are usually used by large population, such that different schools can use such tests to assess their performance with other schools.

Specifically, a standardized test may be said to be tests that are same in marking across all candidates who sat the test using same time and same conditions. It is very important that all contaminating variables are controlled because it will allow for a relative comparison of candidates' performances. Osadebe (2004) describes standardization of tests as the method of creating a consistent test which will involves creation of standards while Meador (2016) viewed standardized tests as tests designed to

suit different situations. He enumerates several benefits that are derivable from standardized tests such as: identifying the fortes and flaws of students with other students of similar grade and level of knowledge, since the results of the tests are public record, it could be used to hold teachers and schools responsible for their students' performances among others.
Kaukab and Mehrunnisa (2016) described standardized tests as the ones that are reliable in marking across all testees, who took the test, given the identical condition and period. Standardized tests are generally considered good tests with three distinct characteristics – reliability, validity and practicality.

Test reliability refers to how consistent and dependable a test is while test validity refers to the extent to which tests scale the concept it was designed to gauge. Practicality relates to how economical, easy to administer and score, and even to interpret. Tests are supposed to be within the means of financial capacity, and easy to administer, score and interpret. Among the three characteristics, practicality may not be considered a fundamental prerequisite. This is because if a test is expensive, and not easy to score or interpret, that may not necessarily affect reliability and validity of the test. Costas (2014) enumerated four basic functions of standardized tests, which include Selection, Classification, Assessment, and Diagnosis.

There are two popular theories and approaches for analyzing test items. They are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT is an old test theory and was established on the suggestion that there will always be measurement error, this error is random, and it is part of the observed score. The basic assumptions of this theory are that there is an error, and the error is a component of a true score and such error is not dependent of the error of other measures. Stage (2004) viewed that CTT's focus mainly on test-level information and does not provide information on each item that make up the test. This is because there are no theoretical models to relate candidates' abilities on each item.

In contrast, IRT is stronger than CTT and based on the probability that candidates' success is at item level rather than test level. According to Fan (1998), IRT focuses mainly on individual item information in contrast to the CTT's primary focus on the total test data. IRT models incorporate groups of postulations. Also, expediency of each postulation depends on the behaviours of the test items and the capability of various expectations about the test items.

Fan (1998) discoursed that though CTT has helped experts in educational evaluation for most of the 20th century, IRT has seen quicker growth in modern periods. Although CTT's major attention is on totality of test data, item analysis in terms of item difficulty and discrimination are equally important aspect of the CTT.

Studies have compared Classical Testing and Item Response Theories and the results are divergent. While some studies supported the superiority and popularity of CTT over IRT, other studies opined that CTT is still of great relevance in assessment., Awopeju and Afolabi (2016) did a comparative analysis of CTT and IRT's item difficulty and discrimination with the ability of testees in the Senior School Certificate Examination (SSCE) in Mathematics. The outcome showed that CTT and IRT were alike in assessing item features of numerical and psychometric tests. They also suggested that both test theories could be used as harmonizing processes in the development of public examinations. In the same vein, Adegoke (2013) using Physics as the subject, investigated how comparable item data created from the models of CTT and the 2-parameter model of IRT. Results showed that item data gotten from both models were rather equivalent. However, item data acquired from IRT 2-parameter model looked steadier than those from CTT. Adegoke further established that for item compilation process, the IRT 2-parameter model led to removal of less items than the CTT model. The result suggested that item writers and test development officers should incorporate the IRT model into their test generation procedure. While Lin (2008) examined the level of parallel test forms that could be gathered with the weighted deviations model using both CTT and IRT methods. The results showed that the CTT approach performed as well as, and even better than the IRT approaches in gathering forms equivalent. Due to this inconsistency and to add to literature in

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

test theories, the present study aimed at constructing and standardizing Economics Achievement Test using the two test theories. Specifically, this study presents the comparative item analysis of the standardization procedure of the Economic Achievement Test (EAT) using classical testing (CTT) and Item response (IRT) methods.

Research Questions

(1) How spread are the items on Table of Specification are the unstandardized Economics Achievement Test (EAT)?

(2) What are the range of the difficulty and discriminating indices of the Economics Achievement Test (EAT)?

(3) Which parameter model best fits the Economics Achievement Test (EAT)?

(4) What are the ranges of the values of parameters of the Economics Achievement Test (EAT)?

Methodology

Research design for this research was a descriptive survey. This method was used to collect data because it enabled researchers to describe systematically the characteristic features of a given population (Nwadinigwe & Azuka-Obieke 2012). The population encompassed all senior secondary school II students in Ogun state. One thousand, two hundred (1,200) students participated in the exercise, 763 from public schools while 437 are from the private schools . Ogun State in the southwest geopolitical zone of Nigeria was purposively selected for this study. Ogun state was selected for its heterogeneous attributes among students in the state. The state consists of urban and rural, private and public schools as well as boys and girls. The multi stage sampling process was adopted for this study. Out of the 20 local government area, six local government areas were randomly selected. From the selected LGAs, eight schools were randomly selected. Four of these schools were privately owned and four were publicly owned. From each selected school, an intact senior secondary two (SSII) Economics class was chosen randomly. The table for the distribution of the sample is as shown in Table 1. Economics is a subject in Business/Humanities and not a compulsory subject.

PUBLIC SCHOOLS				***PRIVATE SCHOOLS***			
SCHOOL	***NO OF STUDENT***	***SCHOOL***	***NO OF STUDENT***	***SCHOOL***	***NO OF STUDENT***	***SCHOOL***	***NO OF STUDENT***
SCH A	73	SCH G	43	SCH A	35	SCH G	43
SCH B	43	SCH H	63	SCH B	43	SCH H	63
SCH C	65	SCH I	68	SCH C	15	SCH I	54
SCH D	95	SCH J	54	SCH D	10	SCH J	34
SCH E	63	SCH K	73	SCH E	63	SCH K	28
SCH F	63	SCH L	60	SCH F	25	SCH L	24
TOTAL			**763**	**TOTAL**			**437**

A 120-Item Economics Achievement Test (EAT) was constructed by the researchers who are experts in test construction, using the teaching curriculum on Economics used in the Nigerian senior secondary schools and examination syllabi by public examination bodies in Nigeria. The Items were reviewed by Economics experts using the Table of Specifications designed from Bloom's taxonomy. The experts were teaching Economics in senior secondary schools and were examiners in the subject. The content validity of the EAT was high and it has a reliability coefficient of .799 using Kuder-Richardson i.e. KR-20 as a measure of internal consistency.

Results And Discussion

Research Question 1: How spread are the items on Table of Specification are the unstandardized Economics Achievement Test (EAT)?

The Table of Specification is as shown in Table 2.

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Table 2:
Table of Specification the Economics Achievement Test (EAT)

Content	Knowledge 28%	Comprehension 23%	Application 20%	Analysis 15%	Synthesis 7%	Evaluation 7%	Total 100%
Meaning and Concept of Economics	8	6	4	4	2	2	**26**
Basic tools of Economic Analysis	6	4	4	2	2	2	**20**
Basic Economics Problems	4	6	4	4	2	2	**22**
Concepts of Demand and Supply	8	4	4	4	2	2	**24**
Business Organization	4	3	3	2	1	1	**14**
Theory of Production	4	3	3	2	1	1	**14**
TOTAL	**34**	**26**	**22**	**18**	**10**	**10**	**120**

From Table 2, the un-standardized Economics Achievement Test (EAT) items were multiple choice formats with four options (one key and three distracters). The items were developed using Bloom's taxonomy with Table of Specification of Knowledge (34%), Comprehension (26%), Application (22%), Analysis (18%) Synthesis, (10%), and Evaluation,(10%). (Topics were spread across senior secondary school one syllabus which covered Meaning and concept of Economics(26%) Basic tools of Economic Analysis -20%, Basic Economic Problems -22%, Concept of Demand and Supply -24%, Business Organizations – 14%, and Theory of Production -14%). This was to ensure that all schools have completed the topics before the test.

Table 2 revealed that both content and cognitive levels were well spread across board. This is an indication that the unstandardized is representative enough. This in support of Asuru (2015)'s view, that Table of Specification helps in showing the correct number of items in the proper content area and guarantees a balance between them thereby avoiding the tendency of developing more questions in some areas while some areas are neglected.

Research Question 2: What are the ranges of the difficulty and discriminating indices of the Economics Achievement Test (EAT)?

Table 3 depicts the difficulty indices of the 120 unstandardized Economics Achievement Test items using CTT.

Table 3:
Difficulty Indices of Un-standardized Draft Economic Achievement Test using CTT

ITEM NO	P*	ITEM NO	P*	ITEM NO	P*	ITEM NO	P*
1	74.22	51	40.17	89	31.8	101	23.42
4	66.7	13	39.79	43	30.76	37	23.33
12	64.82	95	39.6	56	30.29	76	22.67
99	63.22	61	39.51	48	30.2	103	22.58
7	58.23	74	39.32	71	30.2	70	22.48
38	55.69	3	39.13	91	30.01	10	22.11
55	55.13	16	38.95	90	29.92	18	22.11
21	54.84	35	38.85	84	29.26	52	22.11
65	54.19	62	38.85	41	28.22	118	22.11
64	54.09	8	38.76	117	28.03	53	22.01
6	51.46	111	38.76	26	27.47	24	20.32
119	51.18	46	37.44	100	27.47	109	20.32
2	50.99	105	37.06	114	27.47	92	19.76
115	50.61	11	36.78	97	27.28	108	19.66
32	50.33	49	36.41	77	27.09	23	19.59
81	49.58	63	36.31	82	26.81	75	19.00
69	49.48	73	36.31	96	26.43	116	18.44
67	48.82	93	35.18	107	26.34	45	17.31
33	48.07	27	35.09	25	26.25	68	16.93
113	45.06	66	34.9	79	26.25	94	16.56
104	45.06	57	34.24	80	25.31	110	16.46
42	45.06	87	34.24	88	25.31	9	15.73
98	43.56	72	33.58	28	25.21	112	15.71
78	42.43	20	33.21	120	24.46	47	14.77
19	42.14	22	32.93	5	24.18	15	14.02
86	41.77	85	32.83	50	24.18	17	13.73
39	41.77	44	32.74	58	24.18	30	12.42
60	41.49	40	31.89	106	23.99	14	10.82
34	41.11	102	31.89	31	23.71	83	8.56
59	40.55	36	31.8	54	23.42	29	0.00

P=difficulty index

Item difficulty using CTT, measure how easy an item is and denoted as p. It is simply the percentage of testees that got the item correctly. The greater the fraction of those that got an item right, the easier the item, meaning that the higher the difficulty index, the easier the item is. This index ranges from 0 to 100; the higher the value, the easier the question. According to Ado (2015), difficulty indices have interpretations that means p≤.30 are considered difficult; 0.31 ≤0.70 are considered moderately difficult; and p> 0.70 are easy items while some researchers opined that 0.00 – 0.20 is very difficult 0.21 – 0.80 is moderately difficult and 0.81 – 1.00 is very easy.

Using Ado (2015)'s range, it could be seeing from Table 3, that the difficulty indices of 54 items ranged

0< p <.30 which was an indication that these items are difficult; in fact one item was 0.00 meaning that all candidate failed the item. Sixty-five of the items had their difficulty indices ranged between 30.01 and 66.7; these items are moderately difficult. It is an interesting note that only one item of EAT is easy with an index of 74.22. This is an indication that the items on the EAT were well spread across difficulty levels.

Table 4 shows the discriminating indices of the 120 unstandardized Economics Achievement Test items.

Table 4: Discriminating Indices of Unstandardized Economics Achievement Test Using CTT

ITEM NO	r_{pbi}*	ITEM NO	r_{pbi}*	ITEM NO	r_{pbi}*	ITEM NO	r_{pbi}*
67	0.66	33	0.36	2	0.2	71	0.08
86	0.6	79	0.36	10	0.2	82	0.08
64	0.57	11	0.34	37	0.2	53	0.07
98	0.57	35	0.34	40	0.2	75	0.06
32	0.51	62	0.34	100	0.2	22	0.05
115	0.51	13	0.33	101	0.2	118	0.05
8	0.51	44	0.33	56	0.19	25	0.03
73	0.5	102	0.33	96	0.18	66	0.03
4	0.49	60	0.32	74	0.17	15	0.03
55	0.49	114	0.32	105	0.17	20	0.02
3	0.48	21	0.31	39	0.16	27	0.02
6	0.47	120	0.31	117	0.16	80	0.02
61	0.47	78	0.3	36	0.15	106	0.02
95	0.46	84	0.3	77	0.15	108	0.02
99	0.45	24	0.29	116	0.15	23	0.02
93	0.45	18	0.28	110	0.15	14	0.02
119	0.44	113	0.27	1	0.14	17	0.01
65	0.43	107	0.27	68	0.14	109	0
81	0.43	92	0.27	26	0.12	9	0
69	0.41	5	0.26	49	0.12	29	0
46	0.41	16	0.26	51	0.11	83	-0.01
38	0.4	41	0.26	58	0.11	70	-0.03
104	0.4	103	0.26	89	0.11	76	-0.03
87	0.4	85	0.25	111	0.11	88	-0.03
34	0.39	59	0.24	112	0.11	54	-0.07
12	0.38	94	0.24	42	0.1	47	-0.08
63	0.38	72	0.22	43	0.1	45	-0.1
28	0.37	97	0.22	90	0.1	50	-0.11
7	0.36	57	0.21	31	0.09	30	-0.15
19	0.36	91	0.21	48	0.09	52	-0.17

*r_{pbi} = Discriminating index

Item discrimination is the capability of an item to distinguish between higher ability testees and lower ability testees. The crux of item discrimination statistics is to eliminate items that do not behave as expected in the tested group.

Table 4 displays the discriminating indices of EAT items that range from -0.17 to 0.66. Ado (2015) postulated that discriminating index of $r_{pbi} \geq 0.40$ mean that items are functioning acceptably, $0.30 \leq r_{pbi} \leq 0.39$ mean good item with little or no revision, $0.20 \leq r_{pbi} \leq 0.29$ mean that items is marginal and need revision while $r_{pbi} \leq 0.19$ are poor items and need to be eliminated or completely revised. Negative value of discriminating index is an indication that the item should not be used since it cannot distinguish between high achiever and low achievers. Out of the 120 items on EAT, 24 items are functioning satisfactorily, 20 items are good items, 22 items are marginal and could be revisited to remove any ambiguity while 54 items were eliminated. In all, 66 items were fit for the Economic Achievement Test using CTT.

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Research Question 3: What parameter model best fit the Economics Achievement Test (EAT)?

According Adegoke (2013) items acquired from IRT 2-parameter model looked steadier than those from CTT which showed that IRT may provide a better alternative to CTT. IRT model the ability of testees and the probability of responding an item correctly based on the pattern of responses to the items that make up the test. Since IRT makes use of probability, its efficiency is largely dependent on models. There are 3 most prominent models namely: One-Parameter Logistic (1-PL), Two-Parameter Logistic (2-PL) and Three-Parameter (3-PL) Models. In determining the Parameter model that is best to analyse the data that was generated, Akaike Information Criterion (AIC) was done. Akaike Information Criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. Table 4 shows the comparison of AICs the 3 models. The lower the AIC, the better the model fit for the data.

Table 5:
Comparison of AICs of One, Two and Three Parameter Model

	AIC	***AICc***	*SABIC*	*HQ*	*BIC*	*X2*	*df*	*p*
*itemtype = " **1PL** ",*	*151514.2*	***151544.6***	*151734.5*	*151743.0*	*-75636.11*	*NaN*	*NaN*	*NaN*
*itemtype = " **2PL** ",*	*148915.0*	***149050.7***	*149351.9*	*149368.8*	*-74217.48*	*2837.26*	*119*	*0*
	AIC	***AICc***	*SABIC*	*HQ*	*BIC*	*X2*	*df*	*p*
*itemtype = " **2PL** ",*	*148915.0*	***149050.7***	*149351.9*	*149368.8*	*-74217.48*	*NaN*	*NaN*	*NaN*
*itemtype = " **3PL** ”,*	*148790.4*	***149145.5***	*149445.8*	*149471.1*	*-74035.21*	*364.541*	*120*	*0*

From Table 5, comparing AICs of 1PL value of 151544.6 and 2PL value of 149050.7, it could be seen that the 2Pl value is lower than that of the 1PL. Further cursory look at AICs of 2PL value of 149050.7 and 3PL value of 14145.5, it could be seen that the 2PL value is lower than that of the 3PL. Therefore it could be concluded that relatively, 2-Parameter model is best fit for the data. 2-Parameter model only accommodate 'a' and 'b' parameters in the IRT anaylsis.

Research Question 4: What are the ranges of the values of parameters of the Economics Achievement Test (EAT)?

Table 6: 'a' Parameter of Unstandardized Draft Economic Achievement Test

Item No	a	Remark	Item No	a	Remark	Item No	a	Remark
1	0.45	Good	65	0.91	Good	22	0.08	Poor
2	0.48	Good	67	1.82	Good	23	-0.01	Poor
3	0.90	Good	69	1.11	Good	25	0.08	Poor
4	1.54	Good	72	0.31	Good	26	0.20	Poor
5	0.78	Good	73	1.32	Good	27	0.07	Poor
6	1.16	Good	74	0.51	Good	30	-0.40	Poor
7	0.75	Good	77	0.37	Good	31	0.11	Poor
8	1.19	Good	78	0.61	Good	36	0.27	Poor
10	0.51	Good	79	1.11	Good	37	0.30	Poor
11	0.57	Good	81	0.92	Good	43	0.14	Poor

12	1.08	Good	84	0.85	Good	45	-0.29	Poor
13	0.60	Good	85	0.56	Good	47	-0.19	Poor
16	0.58	Good	86	1.55	Good	48	0.29	Poor
18	0.69	Good	87	0.98	Good	50	-0.25	Poor
19	0.75	Good	91	0.45	Good	52	-0.41	Poor
21	0.65	Good	93	1.02	Good	53	0.26	Poor
24	0.87	Good	95	1.04	Good	54	-0.19	Poor
28	1.15	Good	96	0.49	Good	58	0.33	Poor
29	0.38	Good	97	0.44	Good	66	0.15	Poor
32	1.29	Good	98	1.46	Good	68	0.36	Poor
33	0.74	Good	99	1.11	Good	70	-0.03	Poor
34	0.88	Good	100	0.53	Good	71	0.05	Poor
35	0.73	Good	102	0.77	Good	75	0.16	Poor
38	1.07	Good	103	0.85	Good	80	0.02	Poor
39	0.30	Good	104	0.90	Good	82	0.19	Poor
40	0.62	Good	107	0.78	Good	83	-0.10	Poor
41	0.59	Good	113	0.69	Good	88	-0.01	Poor
42	0.19	Good	114	0.80	Good	89	0.10	Poor
44	0.87	Good	115	1.02	Good	90	0.11	Poor
46	0.76	Good	117	0.39	Good	94	0.52	Poor
49	0.24	Good	119	0.99	Good	101	0.30	Poor
51	0.42	Good	120	0.61	Good	106	-0.04	Poor
55	1.20	Good	64	1.38	Good	108	-0.07	Poor
57	0.55	Good	9	0.09	Poor	109	-0.20	Poor
59	0.73	Good	14	0.14	Poor	110	0.50	Poor
60	0.73	Good	15	0.10	Poor	112	0.52	Poor
61	1.10	Good	17	0.10	Poor	116	0.41	Poor
63	0.72	Good	20	0.03	Poor	118	0.04	Poor

Table 7: 'b' Parameter of Un-standardized Draft Economic Achievement Test

Item No	b	Remark	Item No	b	Remark	Item No	b	Remark
1	-2.21	Good	67	0.07	Good	27	9.36	Poor
2	0.04	Good	69	0.08	Good	30	-5.14	Poor
3	0.68	Good	72	2.41	Good	31	11.49	Poor
4	-0.54	Good	73	0.59	Good	36	3.05	Poor
5	1.69	Good	74	1.01	Good	37	4.19	Poor
6	0.00	Good	77	2.88	Good	43	6.13	Poor
7	-0.39	Good	78	0.63	Good	45	-5.66	Poor
8	0.53	Good	79	1.18	Good	47	-9.62	Poor
10	2.69	Good	81	0.10	Good	48	3.13	Poor
11	1.10	Good	84	1.24	Good	50	-4.87	Poor
12	-0.59	Good	85	1.44	Good	52	-3.24	Poor
13	0.83	Good	86	0.33	Good	53	5.16	Poor
16	0.92	Good	87	0.84	Good	54	-6.59	Poor
18	2.05	Good	91	2.07	Good	58	3.63	Poor

19	0.55	Good	93	0.77	Good	66	4.53	Poor
21	-0.22	Good	95	0.54	Good	68	4.63	Poor
24	1.83	Good	96	2.28	Good	70	-39.22	Poor
28	1.21	Good	97	2.44	Good	71	16.33	Poor
29	0.12	Good	98	0.28	Good	75	9.10	Poor
32	0.04	Good	99	-0.51	Good	80	46.16	Poor
33	0.20	Good	100	2.03	Good	82	5.43	Poor
34	0.54	Good	102	1.16	Good	83	-23.51	Poor
35	0.76	Good	103	1.70	Good	88	-80.05	Poor
38	-0.19	Good	104	0.32	Good	89	8.43	Poor
39	1.30	Good	107	1.53	Good	90	8.14	Poor
40	1.40	Good	113	0.40	Good	94	3.33	Poor
41	1.76	Good	114	1.42	Good	101	4.09	Poor
42	1.35	Good	115	0.04	Good	106	-28.74	Poor
44	1.01	Good	117	2.63	Good	108	-21.54	Poor
46	0.82	Good	119	0.01	Good	109	-7.06	Poor
49	2.59	Good	120	2.07	Good	110	3.48	Poor
51	1.11	Good	9	20.03	Poor	112	3.46	Poor
55	-0.15	Good	14	15.90	Poor	116	3.82	Poor
57	1.35	Good	15	17.94	Poor	118	33.06	Poor
59	0.66	Good	17	19.47	Poor			
60	0.60	Good	20	29.36	Poor			
61	0.53	Good	22	9.24	Poor			
63	0.93	Good	23	-174.16	Poor			
64	-0.10	Good	25	13.47	Poor			
65	-0.13	Good	26	5.05	Poor			

IRT software named BILOG was used to analyse the data and the results are depicted in Tables 6 and 7. Since the 2-PL model was the best fit for the data for this study, only parameters “a” and “b” are considered.

Out of the 120 items analyzed, only 114 items of the unstandardized Economics Achievement Test items were analysed. Six of the items were not analysed by the software. The items were item Nos 56,62,76,92.105 and 111. It will be interesting to note that the items that were certified poor in IRT could be revisited for lower level ability examinees. Using IRT, total number of items certified good were 77. This number of items generated was higher than that of the CTT. Using CTT method, total number of items that were certified good were 66 while that of IRT were 77 items. This is in support of Adegoke (2013) who found that two-parameter model of IRT led to remover of lesser items from the test papers than CTT model.

The result discovered that 2-parameter model of IRT was the most suitable for calibration or standardization of the Economics Achievement Test. The content and Bloom's taxonomy were well distributed on the Table of specification. This is a sign that the Economics Achievement Test was calibrated using CTT and IRT. This in support of Asuru (2015)'s view, that TOS helps in showing the appropriate number of items in the proper content area and behaviours and ensures a balance between them thereby avoiding the tendency of developing more questions in some areas while some areas are neglected

Conclusion

The study presented the standardization of Economic Achievement Test from the unstandardized Economics Achievement Test (EAT). EAT was valid and reliable before use. The test items were administered to 1,200 senior secondary two (SS II) students. This standardized test will be good instrument in preparing students for internal and external assessments by teachers and in comparing their performance.

Recommendations

This study recommended that in the construction, standardization and selection of items into test papers, both CTT and IRT approaches should be employed. Also the standardized Economics achievement Test (EAT) would be very useful in assessing students and comparing performance across different schools in Ogun state. The test is for senior school student II who have sufficiently completed their first year in the senior school. The test items would be of great help for the students in preparing them for terminal examinations. The method used for calibration in this study may be useful for other subject.

References

Adegoke, B.A. (2013). Comparison of Item Statistics of Physics Achievement Test using Classical Test and IRT Frameworks. *Journal of Education and Practice.* 4(22), www.iiste.org. ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)

Asuru, V.A. (2015). Measurement and Evaluation in education and psychology. Pearl Publishers International Ltd

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative Analysis of Classical Test Theory and IRT Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *Scientific Journal,* 12(28). p263 doi: 10.19044/esj. 12(28)

Costas, H. (2014). Commercial versus internally developed standardized tests: Lessons from a small regional school. *Journal of Education for Business,* 89(1), 42-48. http://dx.doi.org/10.1080/08832323.2012.740519

Fan, X. (1998). IRT and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58(3), 357-381.*

Hambleton, R. K., & Swaminathan, H. (1985). *IRT: Principles and Applications. New York, NY: Springer*

Kaukab, S.R., & Mehrunnisa, S. (2016). History And Evolution Of Standardized Testing – A Literature Review" *International Journal of Research – Granthaalayah,* 4(5) 126-132.

Kimau, J. (2018) Educational Measurement and Evaluation: Types of Tests and Their Construction https://jeremiahkimaublog.wordpress.com/2015/10/07/educational measurement-and-evaluation-types-of-tests-and-and-their-construction/

Lin, C.J. (2008). Comparisons between Classical Test Theory and IRT in Automated Assembly of Parallel Test Forms. *Journal of Technology, Learning, and Assessment,* 6(8). from http://www.jtla.org

Meador, D.(2016). Examining the Pros and Cons of Standardized Testing. http://teaching.about.com/od/assess/a/Standardized-Testing.htm

Nwadinigwe, I. P., & Azuka-Obieke, U. (2012). The impact of emotional intelligence on academic achievement of senior secondary school students in Lagos, Nigeria. *Journal of Emerging Trends in Educational Research and Policy Studies,* 3(4), 395.

Osadebe, P.U. (2014). Standardization of Test for Assessment and Comparing of Students' Measurement. *International Education Studies* 7(5); ISSN 1913-9020 E-ISSN 1913-9039

Stage, C. (2004). *Classical Test Theory or IRT*: The Swedish Experience. Published in Spanish by Centro de Estudios Públicos, Santiago, Chile. www.cepchile.cl