Psychometric Analysis of Economics Achievement Test Using Item Response Theory

Roseline Ifeoma Ezechukwu[1], Basil Chinecherem, E. Oguguo[2,*], Catherine U. Ene[2] & Clifford O. Ugorji[3]

[1]Department of Educational Psychology, Federal College of Education (Technical), Omoku, Rivers State, Nigeria

[2]Department of Science Education, Faculty of Education, University of Nigeria, Nsukka, Enugu State, Nigeria

[3]Department of Physical Science Education, Imo State University, Owerri, Imo State, Nigeria

*Correspondence: Department of Science Education, Faculty of Education, University of Nigeria, Nsukka, Enugu State, Nigeria. E-mail: basil.oguguo@unn.edu.ng

Received: January 8, 2020 Accepted: February 20, 2020 Online Published: April 6, 2020

doi:10.5430/wje.v10n2p59 URL: https://doi.org/10.5430/wje.v10n2p59

Abstract

This study determined the psychometric properties of the Economics Achievement Test (EAT) using Item Response Theory (IRT). Two popular IRT models namely, one-parameter logistics (1PL) and two-parameter logistics (2PL) models were utilized. The researcher adopted instrumentation research design. Four research questions and two hypotheses were formulated to guide the study. The population size is five thousand, three hundred and sixty-two (5,362) from thirty-seven (37) schools. The sample for the study was 1,180 senior secondary school students (SSS3) drawn using multi-stage sampling procedure. The 1,180 students were stratified according to gender which resulted to 885 females and 295 males. The instrument for the study consisted of 50 multiple-choice test items on the economics achievement test, developed by the researchers. Reliability and validity for each item and for the whole test were established according to the one-parameter and two-parameter logistic models. Research question one was answered using 1PLM, while research questions two and three were answered using 2PLM IRT model. Hypothesis one was tested using t-test analysis of difference between the difficulty parameters estimated using 1PLM and 2PLM while hypothesis two was tested using Chi-square. The finding of the study revealed significant difference between the item difficulties estimated using 1PLM and 2PLM. Also the observed scores of the testees on the test items fit the 1PL 2PL models.

Keywords: psychometric properties, item response theory, one –parameter logistics model, two-parameter logistic model and local independence

1. Introduction

This research determines the psychometric properties of Economics Achievement Test (EAT) for Senior Secondary School Students. Economics is one of the senior secondary school subjects that require assessment to ascertain students' basic knowledge, skills, understanding of the concepts and the nature of economic problems in any society. Economics has been defined variously by many authorities. According to Anyanwuocha (2011) Economics deals with 'what is', 'what was', or what will be', and not 'what ought to be'. He went further to say that economics is not normative, but a positive science which describes the working of the present economic institution as they are and not how they ought to be, but offers suggestions for the solution of economic problems which confront society. Ezechukwu and Amaechi (2016) opined that Economics is the study of human endeavours in respect of production, distribution, exchange and consumption. According to Hansen (2001), economics is one of the few social science subjects that heavily utilize statistical and mathematical models to analyze real-life economic problems.

The relevance of economics as a requirement for technological advancement of a nation cannot be underrated. In Nigeria, Economics came into the secondary school curriculum in 1966 (Obemeata, 1991). The objectives of studying Economics according to Asadu (2001) are:

• to enable students to acquire knowledge for the practical solution of the economic problems of Nigerian societies, developing countries and the world at large.

• to prepare and encourage students to be cautious and effective in the management of scarce resources.

• to equip students with the basic principle of economics necessary for useful living.

• to increase students respect for the dignity of labour and their appreciation to economic, cultural and social values of the society.

The objectives discussed tend to suggest that the study of Economics is a form of learning in which knowledge, skills and habits of a group of people are transferred from one generation to the next through teaching, training or research.

Learning is simply described as a change in behaviour as a result of experience (Maduewesi, 2005). According to Black and William (2009), learning is tied to effective assessment by monitoring students' progress and feeding that information back to students. Assessment is the process of gathering information for the purpose of decision making. It involves the collection of information about an individuals' knowledge, skills, attitude, judgement, interpretation and using the data for taking relevant decisions about the individual, instructional process, curriculum or programme (Anikweze, 2010). According to Nwana (2007) assessment generally refers to making a mental note of the state of an activity or the condition of an object without necessarily carrying out any testing or measurement and thereafter deciding whether and to what extent the activity or object possesses the attribute under study. Frey, Schmitt and Allen (2012) outline the relevance of educational assessment to include determining students' progress and helping educators to reflect on their performance and materials, provoking students' thought and action, encouraging students to ask questions and motivating them to learn. Consequently, to ensure effective teaching and learning of Economics in schools, an achievement test that focuses on attainment on individual items will have better utility than one on students' aggregate scores.

An educational measurement scale that has ratio scale, sample independent attributes and students' ability reported on both item and total instrument levels can be developed with the measurement theory called Item Response Theory (IRT), otherwise known as modern theory. Item response theory is another branch of psychometric theory that may be regarded as roughly synonymous with latent trait theory. It is also referred to as the strong true score theory or modern mental test theory since IRT is the most recent body of theory with stronger assumption than classical theory.

This study was anchored on the modern psychometric method called Item Response Theory. The principles of IRT are based on two basic assumptions. First, a more able person could have greater probability of success on assessment of cognitive items such as economics achievement test items than a less able person. Secondly, any person could always be more likely to do better on an easier item than on a more difficult one. IRT assumes item difficulty and is characterized by influencing item difficulty estimates. It involves a class of mathematical models used to predict examinee performance using item and person characteristics. These models were originally developed for items that are scored dichotomously -that is, correct or incorrect. The concept and method of IRT extend to a wide variety of polychotomous models for all types of psychological variables that are measured by rating scales of various kinds (Vander & Hambleton, 1997).

The seven common models are the Rasch Model (1PL), the two-parameter logistic model (2PL), the three –parameter logistic model (3PL), graded model, nominal model, the partial credit model and the rating scale model. The item format for the first three models is dichotomous, for the last three models, it is polychotomous (Embretson & Reise, 2000). According to De Beer (2004), the first three general IRT models vary in terms of the item characteristics. Each IRT model predicts the probability that a certain person will give a certain response to a certain item. The purpose of these models is to probably explain an examinee's responses to test items in economics based on his/her ability. IRT model assumes that the performance of an examinee can be completely predicted or explained from one or more abilities. IRT models the probability of a correct answer using two logistic functions. The one-parameter logistic (1PL) model attempts to address the probability of a correct answer by allowing each question to have an independent difficulty variable. For instance, one-parameter model allows each question on an achievement test to have an independent difficulty variable.

The two-parameter logistic model allows for different discrimination parameters per item and assumes that the guessing parameter equals 0 (Rizopoulous, 2006). Item discrimination is a measure of how well an item is able to distinguish between examinees who answered the item correctly. When the discrimination index is high it means that the item differentiates discriminates between examinee. The two-parameter logistic model (2PL) allows the slope or discrimination parameter (a) to vary across items instead of being constrained to be equal as in the one parameter logistic or Rasch model. This means that both item difficulty (b) and item discrimination (a) are included in the exponential form of a logistic model. The relative importance of the difference between a person's trait level and

item threshold is determined by the magnitude of the discriminating power of the item (Embretson & Reise, 2000). The constant, 1.7, is added to the model as an adjustment so that the logistic model approximates the normal ogive model (Thissen, Steinberg, & Wainer, 1993). The model assumes that the two parameters (difficulty and discrimination) are necessary for an estimate and valid relationship between the probability of a correct response to an item and the trait level (ability) of an individual. Within the latent trait test model, the internal validity of a test is assessed in terms of the statistical fit of each item to the model.

Fit to the model also implies that item discriminations are uniform and substantial, that is, there are no errors in item scoring. Model-data fit issues are a major concern when applying item response theory (IRT) models to real test data. Model fit is defined as how well the model as a whole explained the data. When a model is over identified, it is expected that model fit will not be perfect; it is therefore necessary to determine the actual degree of model fit, and whether the model fit is statistically acceptable. Ideally, indicators should load only on the specific latent variable identified in the measurement model (Kline, 2010). Nkpone (2001) asserted that in the latent trait models, a fit to the model implies validity that item discriminations are uniform and substantial, and there is no error in terms of scoring. One of the basic assumptions of the application of parametric IRT models is that the model is appropriate for the data. This involves choosing the right model and the evaluating model fit (Edelen & Reeve, 2007). The first consideration when choosing the right model is the number of item response categories. The 1 and 2 IRT models can be used for dichotomous data. The assumption of local independence means that, the probability of an examinee getting item correctly is not affected by the answer given to other items in the test. For example, if the responses to one item structurally constrain the possible answers to other items, then the items are not locally independent. If these assumptions are met, an IRT model can be successfully employed (Courville, 2004). This is one of the hallmark assumptions in IRT, and it makes many things possible. It will also be important for estimating examinee trait levels. Conditional independence provides us with statistically independent probabilities for items.

IRT models are extremely helpful in assessing instrument like Economics achievement test when trying to understand students' abilities by examining their test performance. To ensure that Economics achievement test is fair to all examinees, the instrument should be fair. A test instrument is said to be fair when two groups of equal ability with respect to the construct measured by the test earn the same score on each item of the test. The comparisons between results of subgroups give indication of items that are functioning differently for different groups of students. If the test is not fair or yield different scores from subgroups, for instance gender, it is said to suffer from Differential Item Functioning (DIF). In the view of Meredith, Joyce and Walter (2007) differential item functioning means that individuals of equal ability but from different subgroup e.g., males and females, do not have the same probability of earning the same score. Gender is a broad analytic concept which highlights women's roles and responsibilities in relation to those of men.

Empirically, in a study carried out by Obinne (2008) to examine the psychometric properties of the items of the Biology examinations conducted by the National Examination Council (NECO), and the West African Examination Council (WAEC) using the Item Response Theory (IRT). It was found that the Biology examination items from the two examination bodies were equally reliable and valid. Biology items in the NECO-conducted examination for 2001 were more difficult than those of WAEC of the same year. WEAC items were more prone to guessing than those of NECO items. Obinne's study further reveals that negative difficulty estimates indicate that the items are easy while positive difficulty estimates indicate that the items are hard. The findings of Chong (2013) reveal that difficulty parameter or the threshold parameter value tells how easy or how difficult an item is. The findings which revealed that the items were selected based on the b-value range of -3 to +3 corresponds 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. Baker also 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. Discriminating parameter indicates how well an item discriminate between respondents below and above the item threshold parameter, as in the slope of the item characteristics curves (Reeve & Fayers, 2005). Adedoyin (2010) in his study used chi-square test with probability greater than alpha level of 0.05 significant level to select items that fit model. All the empirical studies reviewed, revolve around the major variables of the current study such as development, validation and item response theory. Therefore, the researchers deems it appropriate to review them in this study as they help in understanding what researchers have done before and the gap between such studies and the present study.

In determining the effectiveness of teaching and learning, equating test scores, developing parallel forms of a given achievement test, for example, an economics achievement test and testing for individual differences in cognitive learning outcomes, it is highly desirable to take into consideration some quantitative information regarding the psychometric properties of each item to be included in the final form of the test. Therefore, the problem of the study

is: What are the psychometric properties of an economics achievement test developed and standardized for senior secondary school economics students using the one-and two parameter logistic models; and the influence of each of the models on the item validity indices?

One of the important concerns in education is to develop a standard measure which can be used to estimate student achievement. Further, the researchers are also concerned about the quality of test items and how examinees answer them. The measurement of cognitive ability has prominently featured in the establishment of the psychology of science in general and the development of measuring instruments in particular. Problems are often encountered during the process of constructing instruments, such as lack of capacity to develop and process measures, and interpret them in a meaningful way. Thus, the development of standard measure for students is becoming more complex. The reliability and validity of assessments have not yet lived up to expectation. The main practical and technical problem with assessment is finding suitable criterion measures to provide predictive validity evidence from learning potential measures. This has implications for establishing quality assessment for students. It is therefore the greatest aim of the researchers to use one-and two parameter logistic models to address the measurement-related problems particularly in the analysis of a cognitive test. Therefore, the problem of this study is to determine the psychometric properties of Economics Achievement Test for Senior Secondary School (EAT) Students using IRT models.

1.1 Purpose of the Study

The general objective of this research is to determine the psychometric properties of Economics Achievement Test (EAT) for Senior Secondary School (EAT) Students Specifically, this study:

1. Estimated the item difficulty indices using one-parameter logistic model (1PL).
2. Estimated the item difficulty indices using two-parameter logistic model (2PL)
3. Estimated the item parameter discrimination indices using the two-parameter logistic model (2PL).

1.2 Research Questions

The following research questions were posed to guide this study;

- What are the estimates of the item difficulty indices using the one-parameter logistic model (1PL)?
- What are the estimates of the item difficulty indices using the two-parameter logistic model (2PL)?
- What are the estimates of the item discrimination indices using the two-parameter logistic model (2PL) IRT models?

1.3 Hypotheses

The following null hypotheses were formulated and tested at 0.05 level of significance:

H_{01}: There is no significant difference between the item difficulty parameters estimated by the 1PLM and those estimated with 2PLM.

H_{02}: There is no significant fit between the items of Economics Achievement test based on one parameter logistic model (1PL) and the two-parameter logistic model, (2PL) IRT models.

2. Method

2.1 Design of the Study

The design of this study is an instrumentation research design. Instrumentation research design, according to Ali (2006), is when the major thrust of the study is geared entirely towards the development and standardization of an instrument whose different psychometric properties have been empirically determined. According to Nworgu (2015), an instrumentation research is a type of design aimed at development and certification of efficacy of an instrument for the measurement of a given behaviour of construct.

2.2 Study Participants

The population of the study consisted of all the senior secondary school students that offered Economics in the 2015/2016 academic year in Imo State Nigeria. The population size is five thousand, three hundred and sixty-two (5,362) from thirty-seven (37) schools.

2.3 Sampling Procedures

The sample for the study was drawn using multi-stage proportionate random sampling technique. At the first stage, two secondary schools were sampled from four Local Government Areas through random sampling with replacement. The number is estimated to be about 1,180 students of SS3 in the eight sampled secondary schools which constituted about 22% of the students. The 1,180 students were stratified according to gender which resulted to 885 females and 295 males.

2.4 Instrument used for Data Collection

The Instrument used for data collection was Economics Achievement Test (EAT) developed by the researchers. It has two parts. Part 'A' seeks information about the personal data of the students in terms of name of the school, local government area and gender. Part B of the instrument consists of fifty items achievement test selected from an initial pool of 100 items constructed to cover the SS3 Economics curriculum. Each item in the EAT tests student's ability to define, recognize, explain, analyse, synthesize and evaluate a problem. To determine the number of test items to be generated from each topic, a table of specifications was constructed and used.

Table 1. Test Blue-Print

Content	Knowing 30%	Understanding 25%	Applying 20%	Analyzing 10%	Evaluating 10%	Creating 5%	total
Topic A 30%	5	4	3	1	1	1	15
Topic B 20%	3	2	2	1	1	1	10
Topic C 25%	4	3	2	1	1	1	12
Topic D 15%	2	2	1	1	0	1	7
Topic E 10%	2	1	1	1	0	1	6
Total 100%	16	12	9	5	3	5	50

Field Work, 2019

The relative weights of each topic in terms of difficulty level, coverage, time spent in teaching such a topic was estimated in percentages. Also the relative weights of the cognitive levels in terms of number of questions were considered in the table of specifications and represented in percentages. The SS3 Economics Curriculum contents such as Topic A: internal trade and international trade, Topic B: Balance of trade payments, Topic C: Economic development, International Economic Organisation, Topic D: Elements of national income were used in developing the test items. The researchers began with an initial collection of 100 items covering all the topics in the Economics Curriculum. Items in each topic were developed based on level of cognitive ability measured by the test. The number of questions outlined was relative to the size of the content as well as the number of tasks implied in the objectives. Table of specifications was used to ensure its content validity.

All the EAT items of a particular content area were placed together in the test. This was an attempt to avoid a situation whereby varying number of items content were thrown together in a random order in the entire test. The instructions on test direction were made clear, complete and concise so that subjects would know what they were expected to do. A conscious effort was made to vary the position of correct options in order to prevent the occurrence of a systematic answer pattern. Items were ordered according to their difficulty levels with easy items appearing first and the most difficult last in order to motivate the subjects into the test and help minimise anxieties and frustrations which tend to accompany an encounter with difficult items at the beginning, the last item on any page ended on that page with no part appearing on the next. Background information including gender, age of subjects was obtained by a section of the test.

2.5 Instrument Validation

The EAT items were submitted to two experienced economics teachers and two experts in Measurement and Evaluation Unit Department of Science education, university of Nigeria, Nsukka. The experts were asked to examine and critique the test items in order to determine the appropriateness of the items for the purpose of the study and also to identify any error, repetition in the items and make suggestions as appropriate. Based on the suggestions of the experts, the items were revised accordingly; some items were modified while others were removed due to irrelevance. In all, fifty (50) items were retained and used for the study.

2.6 Reliability of the Instrument

To establish the reliability of the instrument, EAT was subjected to trial testing. The instrument was administered on a sample of fifty (50) students randomly drawn from schools outside the main study area. The scores obtained from the trial testing were subjected to Kudder-Richardson (KR_{20}) formula to determine the internal consistency of the EAT. The decision to use $K\text{-}_{20}$ was used due to the fact that the items were dichotomously scored for a single administration. The reliability coefficient calculated was found to be 0.89, which was found to be high enough for the study.

2.7 Administration of the Instrument

The instrument was administered to the SS3 students who offered Economics in the sampled schools by the researchers and two research assistants. Guidelines were given to the research assistants to ensure conformity in test administration across the sampled schools. The students were required to indicate the correct option to each question by circling the correct option. The items were scored. A correct answer gets a score of 2 marks while a wrong option got a score of zero marks. The total score was 100 marks (100%) and the least score was 0 mark (0%).

2.8 Data Analysis

Research question 1 was answered using 1PLM, while research questions 2 and 3 were answered using 2PLM IRT model. Hypothesis 1 was tested using t-test analysis of difference between the difficulty parameters estimated using and 1PLM and 2PLM while hypothesis 2 was tested using Chi-square, p-value and -2LL tests of fit.

3. Results

Table 2. Difficulty Parameters of the Items Estimated from 1PLM

Item ID	B	Item ID	B	Item ID	B
ITEM01	-2.233	ITEM19	0.157	ITEM36	0.253
ITEM02	-1.37	ITEM20	-0.039	ITEM37	-0.139
ITEM03	-2.002	ITEM21	1.087	ITEM39	0.44
ITEM04	-2.706	ITEM22	0.81	ITEM40	1.468
ITEM05	0.23	ITEM23	0.06	ITEM41	-0.139
ITEM06	-1.79	ITEM24	0.347	ITEM42	1.37
ITEM08	0.811	ITEM25	-0.459	ITEM43	-0.038
ITEM09	1.477	ITEM26	0.253	ITEM44	0.441
ITEM10	-0.822	ITEM27	-0.039	ITEM45	1.086
ITEM11	0.251	ITEM28	0.993	ITEM46	-0.457
ITEM12	-0.821	ITEM29	0.347	ITEM47	-1.221
ITEM13	0.155	ITEM30	1.086	ITEM48	-1.928
ITEM14	-0.695	ITEM31	-0.038	ITEM49	0.158
ITEM15	0.156	ITEM32	0.625	ITEM50	-0.457
ITEM16	-0.352	ITEM33	0.533		
ITEM17	1.571	ITEM34	1.18		
ITEM18	-0.039	ITEM35	0.44		

Field Work, 2019

Table 2 presented the difficulty indices estimated using 1parameter logistic model (1PLM). The maximum difficulty index is 1.571, while the minimum difficulty index is -2.706 for Item 17 and Item 4 respectively. The table also revealed that items 7 and 38 are flagged off, as a result of having unacceptable difficulty indices.

Table 3. Difficulty Parameters of the Items Estimated from 2PLM

Item ID	B	Item ID	B	Item ID	B
ITEM01	-4	ITEM19	-0.828	ITEM36	-0.484
ITEM02	-4	ITEM20	-0.767	ITEM37	-0.9
ITEM03	-4	ITEM21	0.426	ITEM39	-0.46
ITEM04	-4	ITEM22	0.114	ITEM40	1.156
ITEM05	-0.722	ITEM23	-0.903	ITEM41	-1.134
ITEM06	-3.396	ITEM24	-0.532	ITEM42	0.83
ITEM08	0.112	ITEM25	-1.459	ITEM43	-0.958
ITEM09	0.983	ITEM26	-0.755	ITEM44	-0.327
ITEM10	-1.835	ITEM27	-1.07	ITEM45	0.437
ITEM11	-0.589	ITEM28	0.383	ITEM46	-1.406
ITEM12	-2.262	ITEM29	-0.516	ITEM47	-2.413
ITEM13	-0.892	ITEM30	0.62	ITEM48	-3.51
ITEM14	-1.903	ITEM31	-0.807	ITEM49	-0.843
ITEM15	-0.636	ITEM32	-0.144	ITEM50	-1.799
ITEM16	-1.63	ITEM33	-0.343		
ITEM17	1.338	ITEM34	0.706		
ITEM18	-1.312	ITEM35	-0.457		

Field Work, 2019

Table 3 presented the difficulty indices estimated using two parameter logistic model (2PLM). The maximum difficulty index is 1.338, while the minimum difficulty index is -4.00 for Item 17 and Item 4 respectively. The table also revealed that items 7 and 38 are flagged off, as a result of having unacceptable discrimination and difficulty indices.

Table 4. Distribution of Discrimination Parameters of the Items

Item ID	A	Item ID	A	Item ID	A
ITEM01	0.238	ITEM19	0.378	ITEM36	0.628
ITEM02	0.186	ITEM20	0.651	ITEM37	0.615
ITEM03	0.308	ITEM21	0.503	ITEM39	0.333
ITEM04	0.461	ITEM22	0.396	ITEM40	0.345
ITEM05	0.369	ITEM23	0.416	ITEM41	0.436
ITEM06	0.44	ITEM24	0.397	ITEM42	0.447
ITEM08	0.414	ITEM25	0.479	ITEM43	0.464
ITEM09	0.434	ITEM26	0.336	ITEM44	0.528
ITEM10	0.509	ITEM27	0.399	ITEM45	0.484
ITEM11	0.469	ITEM28	0.381	ITEM46	0.505
ITEM12	0.379	ITEM29	0.414	ITEM47	0.489
ITEM13	0.344	ITEM30	0.298	ITEM48	0.474
ITEM14	0.427	ITEM31	0.6	ITEM49	0.369
ITEM15	0.548	ITEM32	0.393	ITEM50	0.359
ITEM16	0.36	ITEM33	0.297		
ITEM17	0.337	ITEM34	0.343		
ITEM18	0.306	ITEM35	0.336		

Field Work, 2019

Table 4 presented the discrimination indices estimated using two parameter logistic model (2PLM). The maximum discrimination index is 0.651, while the minimum discrimination index is 0.186 for Item 20 and Item 2 respectively. The table also revealed that items 7 and 38 are flagged off, as a result of having unacceptable discrimination indices.

Table 5. t-test of Difference between the Difficulty Parameters of 1PLM and 2PLM

$\bar{X}_1$	$\bar{X}_2$	SD_1	SD_2	t_{Cal}	df	t_{tab}	$p-value$	$Decision$	
0.00	-.977	1.00	1.381	15.70	47	2.00	0.000	Reject	H_{01}

Field Work, 2019

Table 5 presented the result of the t-test analysis of difference between the difficulty parameters estimated using and 1PLM and 2PLM. The calculated t-value (t_{Cal}) is 15.70, which is greater than the tabulated t-value (t_{tab}) of 2.00. The null hypothesis is therefore rejected. Hence, the researchers concluded that at 95% confidence level there is a significant difference between the difficulty indices estimated using 1PLM and those estimated using 2PLM.

Table 6. Overall Model Fit of the Test Items

Test	Items	Chi-square	Df	P- value	-2LL
Full Test	48	1319.998	624	0.000	2627

Field Work, 2019

Table 6 presented the Chi-square, p-value and -2LL tests of fit for the observed scores of the testees on the test items. The p-value of 0.000, which is less than 0.05 and a large -2LL indicated that the observed scores of the testees on the test items does fit the 1PL and 2PL models.

4. Discussion

The result of analysis on the item difficulty estimated with 1PLM and 2PLM revealed that the maximum and minimum difficulty indices are 1.571 and -2.706 respectively for 1PLM and 1.338 and -4.00 respectively for 2PLM. This shows that the range of item difficulty obtained using 1PLM is -2.706 – 1.571, while that of 2PLM is -4.00 – 1.338. These are within the acceptable range of item difficulty as stated in Baker (2001). The finding agrees with (Chong, 2013) that difficulty parameter or the threshold parameter value tells how easy or how difficult an item is. The finding also 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 maximum and minimum discrimination indices are 0.651 and 0.186 respectively. This range is within the acceptable range of discrimination index stated by Baker and Kim (2004). The finding of the study also revealed that there is a significant different between the item difficulty estimated using 1PLM and 2PLM. Also the observed scores of the testees on the test items fit the 1PL and 2PL models. The findings corresponds with Adedoyin (2010) who in his study used chi-square test with probability greater than alpha level of 0.05 significant level to select items that fit model. Nkpone (2001) asserted that in the latent trait models, a fit to the model implies validity that item discriminations are uniform and substantial, and there is no error in terms of scoring.

5. Conclusion

The findings on the range of item difficulty implies that a wide range of item difficulty have been included in the test bank (EAT). This reveals item strata, which identify statistically distinct difficult levels. The information provided will assist teachers in ensuring that items with wide range of difficulties are selected from the EAT during classroom tests or examinations. This will help to improve the teaching process. Also, the test can easily be tailored to proficiency, with easy items for those who show low proficiency and difficult items for those who show higher proficiency. The finding on the range of item discrimination implies that students administered with the test (EAT) may be scored on the same scale, even though they may not respond to the same set of items. The finding will also help teachers to ensure that items with high discrimination powers are included in a test. However, the finding of the study that there is a significant difference between the item difficulty estimated with 1PL and 2PL models implies that the models may be scales themselves (just like temperature values obtained the same object and at the same time using Celsius and Kelvin scales) because it is expected that difficulty estimated using the two models will be the same or at least will not differ significantly.

Acknowledgement

We thank our colleague's and all the authors whose works were consulted during the process of this study.

References

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 November 30, 2012 from http://www.uniBotswana./journal/ education/science

Ali, A. (2006). *Fundamentals of research in education.* AWKA: Meks Publishers (Nig.).

Anikweze, C. M. (2010). *Measurement and evaluation for teacher education.* Enugu: Snaap Press.

Anyanwocha, R. A. I. (2001). *Fundamentals of economics*. Onitsha: Africana FED.

Asadu, I. N. (2001). *Trend in student's enrolment and performance in senior secondary certificate examination in Economics*. Unpublished doctoral dissertation, University of Nigeria, Nsukka.

Baker, F. B. (2001). *The basics of item response theory*. New York, NY: ERIC Clearinghouse.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker. https://doi.org/10.1201/9781482276725

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education, 5*, 9. https://doi.org/10.1186/1472-6920-5-9

Black, P. J., & William, D. (2009). Assessment and classroom learning. *Assessment in education. 5,* 7-74. https://doi.org/10.1080/0969595980050102

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

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*, Ph.D Dissertation. Texas A & M University.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology, 30*(4), 52-58. https://doi.org/10.4102/sajip.v30i4.175

Embertson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ezechukwu, R. I., & Amaechi, C. E. (2016). Development of economics mock examination for senior secondary school students in Imo State. *ASEREN Journal of Educational Research and Development (AJERD*)*, 3*(2), 50-58.

Federal Ministry of Education (2008). *National Policy on Education* (Rev. ed.). Abuja: Federal Government Press.

Hansen, W. L. (2001). Expected proficiencies for undergraduate economics majors. *Journal of Economic Education*, *32*(3), 231-242. https://doi.org/10.1080/00220480109596105

Kline, R. B. (2005). *Principles and practice of structural equation modelling* (2nd ed.). New York: Guilford.

Maduewesi, U. B. (2005). *Curriculum implementation and instruction.* Onitsha: West and Solomon publishing COY LTD.

Meredith, D. G., Joyce, P. G., & Walter, R B. (2007). *Educational research: an introduction* (8th ed.). United State of America: Pearson Press.

Nkpone, H. L. (2001). *Application of latent trait models in the development andstandardization of physics achievement test for senior secondary students*. Unpublished doctoral dissertation, University of Nigeria, Nsukka.

Nwana, O. C. (2007). *Introduction to educational research* (Rev. ed.). Ibadan: HEBN Publishers Plc.

Nworgu, B. G. (2015). *Educational research: basic issues & methodology*. Nsukka: University trust publishers.

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.

Obemeata, J. O. (1991). Pupil's perspective of the purpose of economics education in Nigerian secondary grammar school. *West African Journal of Education, 21*(2).

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire items and scale properties. In P. Fayers and R. D. 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.

Rizopoulous, D. (2006). Ltm: an R package for latent variable modeling and item response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale NJ: Lawrence Erlbaum, Inc.

Vander, L. W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6