World Journal of Innovative Research (WJIR)
ISSN: 2454-8236, Volume-9, Issue-5, November 2020 Pages 130-139

Analysis of Differential Item Functions of Mathematical-Economics Item Structured Test Forms Based on Item Difficulty Levels in Colleges of Education, Enugu State, Nigeria

Salihu Abdullahi GALLE, GBANDE Samson Samuel Kale, IBRAHIM Suleiman Agahu

***Abstract*— The study of Mathematical-economics is paramount to all the Economics students in Nigerian tertiary institutions in the enhancement of growth and nation development. In actualizing the dream goal of the subject two modern theories were anchored thus: Classical Test Theory (CTT) and Item Response Theory (IRT) to evaluate the study. Hence, the researchers evaluated differential item functions of Mathematical-economics item structured test forms (easy-to-hard, hard-to-easy and random versions) based on item difficulty levels in colleges of education, Enugu State, Nigeria. Three research questions guided the study and co-relational research design was adopted. The population of study consisted of all 6420 Federal, State, and Private NCE 1 Economics students 2018-2019 academic season and a simple random sample of 610 NCE1 Economics students were selected for the study. Mathematical-Economics Achievement Test (MEAT) was used for data collection. MEAT was validated which yielded 0.86 validity index and 0.84 reliability index. Data were analyzed using factor analysis (TID, MH, LR CTT-based andX2 and Raju's IRT-based). The result reveals that examinees gain advantages of easy-to-hard than hard-easy or even a random version test forms and differential item structured correct responses to the items for examinees with the same ability levels. It was recommended that economics lecturers in colleges of education should use easy-to-hard test form to evaluate examinees considering with psychometric functions of evaluation for valid decisions.**

***Index Terms*— Mathematical-economics, item structured, test forms, item difficulty, colleges of education.**

I. INTRODUCTION

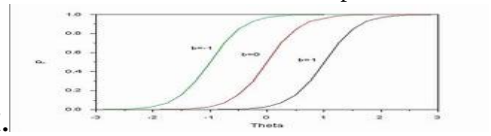
Differential Item Functioning (DIF) has been increasingly applied in fairness studies in psychometric circles. Judicious application of this methodology by the researchers, however, requires an understanding of the technical complexities involved. Hence, high-quality evaluation system because a well structured test forms provides vital feedback to stakeholders and educators regarding students' educational achievement or progress. There is a growing body of literature that addresses the importance of validity, reliability and other quality indicators of assessments [1]. An effective testing instrument is a test that satisfies certain requirements or properties of a useful instrument as judged by experts in educational measurement and evaluation to check for comprehensiveness, adequacy and relevance of the items in agreement with operational chart (Table of specification). [2] View a good measuring instrument should be valid, reliable and usable arrange in the order crucial importance.

According to [3] test is an instrument or procedures which is designed to measure the knowledge, intelligence, ability, traits, skills, aptitude, interest, attitude which an individual or thing exhibits. It is a systematic procedure for observing an individual's behaviour as well as describing such behaviour or performance by numerical scale or category. Educational tests are frequently used to explore individual academic performances, educational needs, and curriculum assessment. Results that are obtained from these tests form the basis for critical decisions to get to know individuals, to employ or place them in institutions or schools, and to select, guide and assess people. As a result, it is essential to prove empirically that test scores have high validity and reliability. What is more, ongoing decisions taken by individual or organizational test developers, practitioners, and interpreters according to test scores depend on developing and implementing eligible methods to examine test development and psychometric qualifications [4].

Test of knowledge and examinations at all stages of education, especially at the higher education level, have been considered an important and powerful tool for decision-making in our competitive society, with people of all ages being evaluated with respect to their achievement, skills and abilities. [5] posited that "the era in which we live is a test-conscious age in which the lives of many people are not only greatly influenced, but are also determined by their test performance." Students consistently perceive test/examination as a source of increased anxiety and a situation engulfed with uncertainty/unfairness in letting them demonstrate their true achievement [6]. The academic results of the Economics departments from Colleges of Education in Enugu State revealed Economics students inability to score high in the subject. There has been mass failure rate in Mathematical-Economics tests/exam. For instance, the results of NCE 1 students for the period of 2016-2019 academic years, 20.23% passed the course while 79.77% failed (Carry Over) [7]. Other factor leading to students failure are pressure of scoring low/high in tests, fear of passing/failing a course, environment of the examination hall

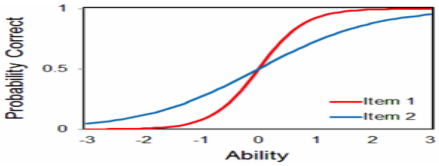
Salihu Abdullahi GALLE, Department of Educational Foundations, Nasarawa State University Keffi, Nigeria.
GBANDE Samson Samuel Kale, Department of Economics, College of Education Akwanga, Nasarawa State, Nigeria
IBRAHIM Suleiman Agahu, Department of Economics, College of Education Akwanga, Nasarawa State, Nigeria

and lack of clarity in instruction for students [8]. These factors or errors could emanate in three categories: the first ''errors inherent in the instrument, ''errors in the use of the instrument and ''errors emanating from the responses of test takers' [9].As examinee's abilities vary their position on the latent construct's continuum changes and is determined by the sample of respondents and item parameters. An item must be sensitive enough to rate the respondents within the suggested unobservable continuum in Graphic 1.



Graphic 1.

Item Difficulty (bi) is the parameter that determines the manner of which the item behaves along the ability scale. It is determined at the point of median probability i.e. the ability at which 50% of respondents endorse the correct answer. On an item characteristic curve, items that are difficult to endorse are shifted to the right of the scale, indicating the higher ability of the respondents who endorse it correctly, while those, which are easier, are more shifted to the left of the ability scale in Graphic 2.



Graphic 2

Item Discrimination (a_i) determines the rate at which the probability of endorsing a correct item changes given ability levels. This parameter is imperative in differentiating between individuals possessing similar levels of the latent construct of interest. The ultimate purpose, for designing a precise measure is to include, items with high discrimination, in order to be able to map individuals along the continuum of the latent trait. On the other hand, researchers should exercise caution if an item is observed to have a negative discrimination because the probability of endorsing the correct answer shouldn't decrease as the respondent's ability increases. Hence, revision of these items should be carried out. The scale for item discrimination, theoretically, ranges from -∞ to +∞; and usually doesn't exceed 2; therefore realistically it ranges between (0, 2). Guessing (c_i) Item guessing is the third parameter that accounts for guessing on an item. It restricts the probability of endorsing the correct response as the ability approaches -∞ [10], [11).

This study anchored on two modern measurement theories, which are the Classical Test Theory (CTT) and the Item Response Theory (IRT) cited in [12]. These two theories are based on different assumptions and use different statistical approaches. CTT is regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in ability of interest. The central model of the CTT is that observed test scores (X) are composed of a true score (T) and an error score (e) where the true and the error scores are independent. The variables are established [13] and best illustrated in the formula: X = T + e. Based on the premise that observed scores are a function of only factors – true scores and measurement error – the theoretical basis for CTT resides in the following formula: X = T + e. This equation represents the three components as discussed above, with T being the hypothetical indicator, X the observed indicator, and e' the amount of disagreement between T and X. IRT is generally regarded as an improvement over CTT. For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information.

For test items that are dichotomously scored, there are three IRT models, known as one Parameter Logistic Model (1PLM), two Parameter Logistic Model (2PLM)and three Parameter Logistic Model (3PLM). A primary distinction among the models is the number of parameters used to describe items. The equation of the Item Characteristics Curve (ICC) for IRT models thus:

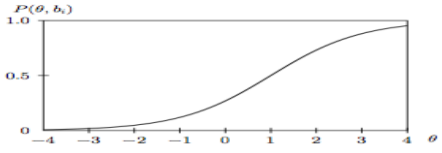
$$1PLM: P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \quad \ldots\ldots\ldots\ldots eqnn(1)$$

$$2PLM: P_i(\theta) = \frac{1}{1 + e^{-D a_i (\theta - b_i)}} \quad \ldots\ldots\ldots\ldots eqn(2)$$

$$3PLM: P_i(\theta) = C_i + (1 - C_i)\frac{1}{1 + e^{-D a_i (\theta - b_i)}} \quad \ldots eqn(2)$$

Where: Pi (θ) is the probability of a current response for the i^{th} item; b_i is the difficulty parameter for the i^{th} item; a_i is the discrimination parameter for the i^{th} item; ci is the guessing parameter for the i^{th} item; θ is the ability level; D represents a scaling factor. These theories enable the studying of tests by identifying parameters of item difficulty, item discrimination and the ability of test takers. CTT and IRT analyze items qualitatively, in terms of their content and form, which includes content validity, as well as item-writing procedures and quantitatively, in terms of their statistical properties, which includes the measurement of item difficulty and discrimination.IRT Assumptions are:

1. Monotonicity – The assumption indicates that as the trait level is increasing, the probability of a correct response also increases,
2. Unidimensionality – The model assumes that there is one dominant latent trait being measured and that this trait is the driving force for the responses observed for each item in the measure,
3. Local Independence – Responses given to the separate items in a test are mutually independent given a certain level of ability,
4. Invariance – We are allowed to estimate the item parameters from any position on the item response curve. Accordingly, we can estimate the parameters of an item from any group of subjects who have answered the item. If the assumptions hold, the differences in observing correct responses between respondents will be due to variation in their latent trait. Item Response Function and Item Characteristic Curve (ICC) in Graphic 3.



Graphic 3

IRT models predict respondents' answers to an instrument's items based on their position on the latent trait continuum and the items' characteristics, also known as parameters. Item response function characterizes this association. The underlying assumption is that every response to an item on an instrument provides some inclination about the individual's level of the latent trait or ability. The ability of the person (θ) in simple terms is the probability of endorsing the correct answer for that item. As such, the higher the individual's ability, the higher is the probability of a correct response. This relationship can be depicted graphically and it's known as the Item Characteristic Curve. As is shown in the figure, the curve is S-shaped (Sigmoid/Ogive). The probability of endorsing a correct response monotonically increases as the ability of the respondent becomes higher. It is to be noted that theoretically, ability (θ) ranges from -∞ to +∞, however in applications, it usually ranges between -3 and + 3.

Consequently, in the literature survey, it was discovered that researchers are not unanimous in their findings as to whether or not altering item structured test forms would affect performance adversely.[14] investigated the invariance properties of one, two and three parameter logistic item response theory models. It examined the best fit among one parameter logistic (1PL), two-parameter logistic (2PL) and three-parameter logistic (3PL) IRT models. The research findings were that two parameter model IRT item difficulty and discrimination parameter estimates exhibited invariance property consistently across different samples and that 2-parameter model was suitable for all samples of examinees unlike one-parameter model and 3-parameter model. Margaret & [15] investigated the effect of test item arrangement on performance in Mathematics among Junior Secondary School Students in Obio-Akpor L.G.A of Rivers State. The findings of the study were that item arrangement based on ascending order of difficulty has a positive and significant effect on students' performance in mathematics at 0.05 alpha level respectively while item arrangement based on descending order has a positive but insignificant effect on student' performance in mathematics.

[16] Examining differential item functions of different item ordered test forms according to Item Difficulty Levels. The finding reveals that item order differentiates the probability of correct response to the items for those at the same ability levels. A test form of sequential easy-to-hard questions brings more advantages than that of a hard-to-easy sequence or a random version. One of such factors is the influence of "test format". Whether test constructors use "multiple-choice", "true-false",' open-ended" or other testing formats in their tests, may influence the test takers' performance [17]). [18] Argued, since none of the test formats is perfect to function well in every context, test constructors must first look into the characteristics of each test format and then make the best selection. With regard to the test format, most of the studies focused on the two commonly-used forms: Open-ended and Multiple-choice forms. [19] Conducted a study on item sequence on test performance. The results of the study revealed that the sequence of items affect foreign language learners' performance. That is, those taking easy to difficult test outperforming students taking the difficult to easy test. The study also bears a set of implications.

[20] Investigated the effects of changing an "easy-to-hard" arrangement to either hard-to-easy or a random arrangement. He found out that the hard-to-easy arrangement was significantly more difficult than the original easy-to-hard order while the random arrangement was not significantly different. [21] Asserted that tiny changes in test format (or arrangement) can make a large difference in student performance. [22] Also found no significant differences between easy-to-hard and hard-to-easy arrangement, easy-to-hard and random order; and hard-to-easy and random order. The main objective of this study focused on the analysis of differential item functions of mathematical-economics item structured test forms based on item difficulty levels in Nasarawa State Colleges of Education

Research Questions

The following research questions guided the study:

1. What are the difficulty levels of test item forms (easy-to-hard and hard-to-easy versions) on examinees based on CTT and IRT?
2. What are the difficulty levels of test item forms (easy-to-hard and random versions) on examinees based on CTT and IRT?
3. What are the difficulty levels of test item forms (hard -to-easy and random versions) on examinees based on CTT and IRT?

II. MATERIAL AND METHODS DESIGN

Design

The researchers adopted a co-relational research design. This is because it involves the collection of differential data form from a group or a random sample of a targeted population [23]. The rational for the design was to evaluate the structured test items forms in agreement with to item difficulty based on Classical Test Theory (CTT) and Item Response Theory (IRT) methods.

Population and Sample

The population of study consisted of all 6420 NCE 1 Economics students in Federal, State and Private Colleges of Education in Enugu State 2018-2019 academic season and a simple random sample of 610 NCE1 Economics students were selected from 3 colleges of education (Federal College of Education, Eha-Amufu [FCOE], State College of Education Osisatech [SCOE] and Private Peace-land College of Education [PCOE]) in Enugu State for this study.300 Economics students from FCOE, 150 Economics students from SCOE and 160 Economics students from PCOE [24].

Before the selection of sample size, lottery method of simple random sampling was employed to selected sample size of 610 NCE 1 Economics students from three colleges of education (Federal College of Education, Eha-Amufu

[FCOE], State College of Education Osisatech [SCOE] and Private Peace-land College of Education [PCOE]) in Enugu State. Serial numbers of the elements on pieces of papers folded and mixed thoroughly before the respondents were asked to pick at once without replacement. This technique gave equal opportunity to the respondents thereby reducing the bias effect that may interfere with the validity and reliability of the study. The three forms of tests thus: Easy-to-Hard (EH), Hard-Easy (HE) and Random Version (RV). The distribution of the NCE1 Economics students according to schools and the test forms is presented in below Table 1.

Table 1: Distribution of the NCE1 Economics Students According to Schools and Test Forms

Colleges of Education (COE) in Enugu State	Tests Forms			Total
	EH	HE	RV	
Federal (FCOE)	100	100	100	300
State (SCOE)	50	50	50	150
Private (PCOE)	53	53	54	160
Total	**203**	**203**	**204**	**610**

Source: *ESCOEED,(2019)*

Instrument for Data Collection

Mathematical-Economics Achievement Test (MEAT) was used as the instrument for data collection. The researchers developed the items after the item analysis of the multiple choice items prepared in the first semester course (ECO 111: Mathematical-Economics) 2018-2019 academic session. According to the item analysis, items with a degree of discrimination of more than 0.30 were selected in such a way that they would not prejudice the validity of the test. A 40 items multiple choice items contained in the Mathematical-Economics Achievement Test was reviewed by the lecturers that were teaching Mathematical-Economics for content validity.

Validity and Reliability of Instrument

Mathematical-Economics Achievement Test (MEAT) was developed by the researchers and subjected to experts' judgment for face and content validation. This was determined through the judgment of four experts, who are knowledgeable in the skills being measured, by checking for appropriateness, comprehensiveness and relevance of the items, clarity of expression and size of print. Two Economics lecturers that were teaching mathematical-economics and two experts' in educational measurement and evaluation that are knowledgeable in the subject from Nasarawa State University Keffi validated the instruments (MEAT). Items that did not measure what they ought to measure were deleted or modified, while good items were retained. The experts verified if the items were in line with the content and objectives stated in the curriculum. The consensus of the experts' judgment yielded 0.86 validity index. The Kuder-Richardson (KR-21) formula was used to determined reliability of the internal consistency of the Mathematical-Economics Achievement Test (MEAT) for the study. Pilot study was conducted on small portion of the population (30 economics students) who are not part of the sample of this study; result for MEAT gave 0.88 reliability index and 0.84 reliability index.

The reliability results of MEAT was compared with the guidelines for interpreting alpha coefficients suggested [25] that "$\alpha \geq 0.9$ excellent, ≥ 0.8 good, ≥ 0.7 acceptable, ≥ 0.6 questionable, ≥ 0.5 poor, ≤ 0.5unacceptable". Therefore, the results of the reliability enabled the researchers to use the instrument (MEAT) for this study, since the correlation was considered high and significant. Therefore, three difficulty levels of the structured tests forms: easy, moderate and hard are presented according to the magnitudes of item difficulty indices in Table 2.

Table 2: Mathematical-Economics Achievement Test Item Difficulty Indices

SN	Items Number	Difficulty Level	Sig
1	1	Easy	0.88
2	2	Easy	0.70
3	3	Easy	0.67
4	4	Easy	0.71
5	10	Easy	0.65
6	15	Easy	0.64
7	8	Easy	0.68
8	20	Easy	0.82
9	25	Easy	0.64
10	30	Easy	0.80
11	6	Easy	0.64
12	19	Easy	0.91
13	31	Easy	0.73
14	22	Easy	0.69
15	38	Hard	0.19
16	36	Hard	0.26
17	33	Hard	0.31

18	35	Hard	0.31
19	5	Hard	0.33
20	27	Hard	0.36
21	40	Hard	0.15
22	24	Hard	0.26
23	17	Hard	0.31
24	39	Hard	0.27
25	7	Hard	0.18
26	34	Hard	0.14
27	23	Hard	0.39
28	37	Moderate	0.43
29	11	Moderate	0.45
30	12	Moderate	0.46
31	13	Moderate	0.53
32	14	Moderate	0.54
33	32	Moderate	0.56
34	26	Moderate	0.43
35	16	Moderate	0.45
36	21	Moderate	0.46
37	9	Moderate	0.53
38	28	Moderate	0.51
39	29	Moderate	0.54
40	18	Moderate	0.49

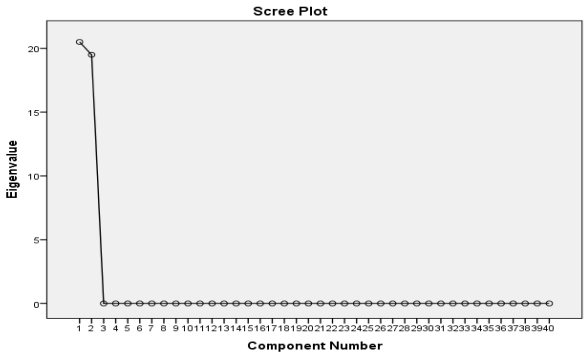
Source: *Researchers Field Work*

Table 2 displayed difficulty indices of the structured test items forms range from 0.14 to 0.91 based on Item Difficulty Index (IDI). When sig (r) ranging from 0 to 1of the correct response in a particular group. This implies that an item with r-value of IDI closer to 0 means the item is hard meanwhile, r-value closer to 1 signifies that the item is easy. According to [26] identified the classification of r-value on Item Difficulty Levels (IDL) for drawing inferences to make decision for valid judgment as following: r-value of 0.00 to 0.39 = Hard, 0.40 to 0.59 = Moderate and 0.60 to 1.00 = Easy. The researchers made used of three tests forms thus: Easy-to-Hard (EH=203), Hard-Easy (HE=203) and Random Version (RV=204).

Procedure for Data Collection and Analysis

The researchers explored whether the data collected would meet the assumptions of IRT for the analyses based on the theory and the type of Parameter Logistic Models (PLM) to be employed in the estimation of item parameters was decided. As a result of the model-data fit analyses, when the number of parameters was 40 items under one Parameter Logistic Models (1PLM), the −2 Loglikelihood value was found to be 7423.2618. Hence, 40 parameters were produced in which solely item hardies were considered for each item as the achievement test consisted of 40 items. Under 2PLM, the −2Loglikelihood value decreased to 7221.6479. The decrease was significant for 40 degrees of freedom (*df*) in the Chi-square (X^2) critical value. Thought, tow Parameter Logistic Models (2PLM) is a model that takes differentiation into account, as well as item discrimination, a total of 80 parameters (40 difficulty and 40 discrimination parameters) were produced for 40 items. For three Parameter Logistic Models (3PLM), the -2 Log likelihood value was 7012.0173 meanwhile, a decrease in the value was not significant.

Consequently, 2PLM and 3PLM is a model that takes guessing parameters into account. Nevertheless, a changed from 2PLMdown to 3PLM result to insignificant decrease in the −2 Log likelihood value led to a decision to apply 2PLM in the estimation of item parameters. Ex-traction method of Principal Component Analysis (PCA) under Factor Analysis (FA) was used to test whether the Mathematical-Economics Achievement Test (MEAT) had a unidimensional. Another assumption of IRT was hold. Data were obtained binary or dichotomously (wrong or right) from the test ranged on a scale mark of 1–0, factor analysis. The gradual trailing off (scree) which was examined to decide the number of factors in PCA is presented in Graphic 1. On test, Graphic 1 clearly reveals that there is a dominant factor. This case shows that unidimensionality assumption of IRT is hold in Graphic 4.



Graphic 4.

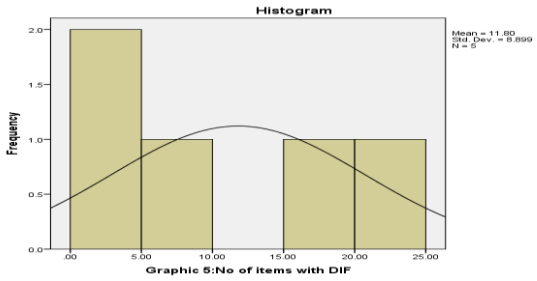
Factor analysis based on IRT-based and CTT-based methods are used in the study for DIF detection: Transformed Item Difficulty (TID), Mantel-Haenszel (MH) and Logistic Regression (LR) CTT-based and Chi-square and Raju's Area IRT-based. In the analyses based on IRT, estimates were made according to 2PLM. BILOG-MG 3 was used in detecting DIF with both CTT and IRT methods. The classification recommended by Educational Testing Service is widely recognized and employed in the field to objectively interpret DIF levels. The following are generally defined DIF

levels although there could be certain changes when specific restrictions of methods are considered [27]: A: Acceptable DIF, B: Moderate DIF and C: High DIF

III. RESULTS

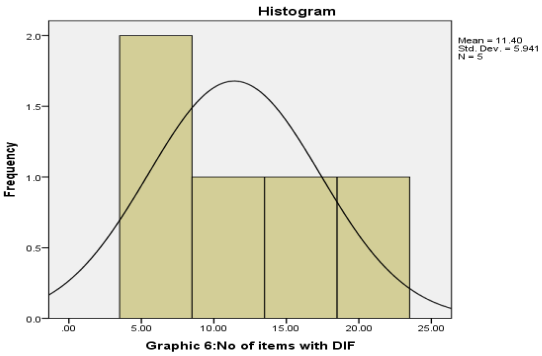
Research question one: What are the difficulty levels of test item forms (easy-to-hard and hard-to-easy versions) on examinees based on CTT and IRT?

Table 3 in Appendix A[1]: shows four items (Items 7, 13, 26 and 40) with significant DIF (Level B and C) based on CTT-based using two methods. While the IRT-based methods, that number increases to twenty items (Items 1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 21, 22, 24, 25, 26, 28, 30, 31, 34, and 40) DIF. Level of DIF of the examinees test was displayed by items 7, 13, 26 and 40based on the two theories (CTT and IRT). It was also observed that the examinees that had EH form of structured test gained more advantaged especially in the first ten test items than their counterpart examinees in the HE forms. And again, this shows that the examinees in the group that were exposed to HE test forms faced disadvantaged of structured test forms. Distribution of number of items with DIF is shown in Graphic 5.



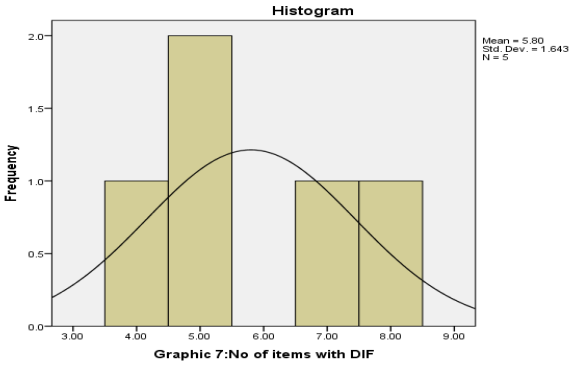
Research question two: What are the difficulty levels of test item forms (easy-to-hard and random versions) on examinees based on CTT and IRT?

Table 4 in Appendix A[2]: shows eight items (Items 4, 5, 10, 15, 16, 17, 20 and 24) with significant DIF (Level B and C) based on CTT-based using two methods. While the IRT-based methods, that number decreases to seventeen items (Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 27, 29, 32, 33, 35, 36, 38, and 39) DIF. Level of DIF of the examinees test was displayed by item 4, 5, 10, 15, 16, 17, 20 and 24 based on the two theories (CTT and IRT). It was also observed that the examinees that had EH form of structured test gained more advantaged especially in the first ten test items than their counterpart examinees in the RV test forms. And again, this shows that the examinees in the group that were exposed to RV test forms faced disadvantaged of structured test forms. Distribution of number of items with DIF is shown in Graphic 6.



Research question three*:* What are the difficulty levels of test item forms (hard -to-easy and random versions) on examinees based on CTT and IRT*?*

Table 5 in Appendix A[3]: shows two items (Items 6 and 18) with significant DIF (Level B and C) based on CTT-based using two methods. While the IRT-based methods that number decreases to eight items (Items 6, 12, 16,22, 28, 34, 37 and 40) DIF. Level of DIF of the examinees test was displayed by item 6 and 18 based on the two theories (CTT and IRT). It was also observed that the examinees that had HE form of structured test gained more advantaged than their counterpart examinees in the RV test forms. And again, the result shows that the examinees in the group that were exposed to RV test forms faced disadvantaged of structured test forms. Distribution of number of items with DIF is shown in Graphic 7.



IV. DISCUSSION

Findings of this study in Table 3in in Appendix A[1] shows four items with significant DIF (Level B and C) based on CTT-based using two methods while the IRT-based methods numbers increases to twenty items DIF. Level of DIF of the examinees test was displayed by four items based on the two theories (CTT and IRT). It was also observed that the examinees that had Easy- Hard (EH) items form of structured test gained more advantaged especially in the first ten test items than their counterpart examinees in the Hard-Easy (HE) forms. And again, this shows that the examinees in the group that were exposed to Hard-Easy (HE) test forms faced disadvantaged of structured test forms. This answered the research question one that states ‘‘*what are the* difficulty levels of test item forms (easy-to-hard and hard-to-easy versions) on examinees based on CTT and IRT?’’ The finding of this study is in agreement with the findings of [28] reveals that item order differentiates the probability of correct response to the items for those at the same ability levels and a

test form of sequential easy-to-hard questions bring more advantages than that of a hard-to-easy sequence or a random version. The findings show that it is essential to arrange tests that are employed to make decisions about people in consideration with psychometric principles.

Furthermore, Table 4 in Appendix A^2 shows eight items with significant DIF (Level B and C) based on CTT-based using two methods. While the IRT-based methods, that number increases to seventeen items DIF. Level of DIF of the examinees test was displayed by eight items based on the two theories (CTT and IRT). It was also observed that the examinees that had EH form of structured test gained more advantaged especially in the first ten test items than their counterpart examinees in the RV test forms. And again, this shows that the examinees in the group that were exposed to RV test forms faced disadvantaged of structured test forms. This answered the research question two that states ''what are the difficulty levels of test item forms (easy-to-hard and random versions) on examinees based on CTT and IRT?''. The finding of this study is in agreement with the findings of [29] revealed that the sequence of items affect foreign language learners' performance. That is, those taking easy to difficult test outperforming students taking the difficult to easy test. The study also bears a set of implications.

Lastly, Table 5 in Appendix A^3 shows two items with significant DIF (Level B and C) based on CTT-based using two methods. While the IRT-based methods that number decreases to eight items DIF. Level of DIF of the examinees test was displayed by two based on the two theories (CTT and IRT). It was also observed that the examinees that had HE form of structured test gained more advantaged than their counterpart examinees in the RV test forms. And again, the result shows that the examinees in the group that were exposed to RV test forms faced disadvantaged of structured test forms. The finding of this study is in agreement with the findings of [30] revealed that item arrangement based on ascending order of difficulty has a positive and significant effect on students' performance in mathematics at 0.05 alpha level respectively while item arrangement based on descending order has a positive but insignificant effect on student' performance in mathematics. Finally, item arrangement based on no particular order of difficulty has a positive and significant effect on students' performance.

V. CONCLUSION

This study was design to evaluate differential item functions of mathematical-economics item structured test forms based on item difficulty levels in Enugu State Colleges of Education. It was concluded that educational and psychological tests should not be influenced by any qualities except examinee abilities and they should remain unbiased without advantageous or disadvantageous on the feedback any groups of examinees. There for it is very essential for test developer to adhering to the basic principles of test and measurement in any forms of test practices in tertiary institutions.

REFERENCES

[1] M. J.Rudolph, K. K. Daugherty, R. M. Elizabeth, V. P. Shuford, L. Lebovitz, & M. V. DiVall. Best practices on exam item construction and post-hoc review. American Journal of Pharmaceutical Education. 2(4),7-20. 2019.
[2] C. M. Anikweze. Measurement and evaluation for teachereducation, (2nd Ed.) Enugu, SNAAP Press, 2016.
[3] Opara, I. M. Psychological Testing, Principles and Techniques. Owerri Career publishers. 2014.
[4] G. Camilli & L. A. Shepard. Methods for identifying biased test items. London, UK: Sage, 1994 In Ö. Çokluk, E. Gül, & C. Doğan-Gül. Examining Differential Item Functions of Different Item Ordered Test Forms According to Item Difficulty Levels.Journal of educational sciences: theory & practice (JESTP-DOI) 16(1), 319-330. 2016.
[5] S. A. Galle & I. J. Kukwi. Effects of Formative Assessment on Econometric Test Anxiety and Students Academic Achievement in Nasarawa State University, Keffi, Nigeria." IOSR Journal of Research & Method in Education (IOSR-JRME), 10(4), 27-36. 2020.
[6] C.D. Spielberger. Test anxiety inventory. Palo Alto, CA: Consulting Psychologists Press, 1985 In S. A. Galle & , I. J. Kukwi. Effects of Formative Assessment on Econometric Test Anxiety and Students Academic Achievement in Nasarawa State University, Keffi, Nigeria." IOSR Journal of Research & Method in Education (IOSR-JRME), 10(4), 27-36. 2020.
[7] Exams and Record, Department of Economics: Federal, State and Private Colleges of Education, Enugu State Enugu State, Nigeria, 2019.
[8] S. A. Galle & I. J. Kukwi. Effects of Formative Assessment on Econometric Test Anxiety and Students Academic Achievement in Nasarawa State University, Keffi, Nigeria." IOSR Journal of Research & Method in Education (IOSR-JRME), 10(4), 27-36. 2020.
[9] C.M. Anikweze. Measurement and evaluation for teachereducation, (2nd Ed.) Enugu, SNAAP Press, 2016.
[10] C. DeMars. Item Response Theory. Cary, NC, USA: Oxford University Press, USA; 2010.
[11] S.E. Embretson & P.R. Steven. Item response theory. Psychology Press, 2013.
[12] S.A. Galle, V. Paul & G. A. Andzutsi. Effect of Changes in Item-Sequence on Students Academic Achievement In Multiple-Choice Test Of Mathematical-Economics In Colleges Of Education, Lagos State Nigeria. World Journal of Innovative Research (WJIR) ISSN: 2454-8236, Volume-9, Issue-4, 69-76, 2020.
[13] M.R. Novick. The axioms and principal results of classical test theory. J. Math. Psych., (1966). 3(2), 1-18 In O. A. Awopeju, E. R. I. Afolabi & O. A. Opesemowo. Investigated the invariance properties of one, two and three parameter logistic item response theory models. Bulgarian Journal of Science and Education Policy (BJSEP), 2(11), 197-219.2019.
[14] O. A. Awopeju, E. R. I. Afolabi & O. A. Opesemowo. Investigated the invariance properties of one, two and three parameter logistic item response theory models. Bulgarian Journal of Science and Education Policy (BJSEP), 2(11), 197-219.2017.
[15] O.I. Margaret & U.I. Victor. Effect of Test Item Arrangement On Performance In Mathematics Among Junior Secondary School Students In Obio/Akpor Local Government Area Of Rivers State Nigeria. British Journal of Education (2017), 8(5), 1-9.
[16] Ö. Çokluk, E. Gül, & C. Doğan-Gül. Examining Differential Item Functions of Different Item Ordered Test Forms According to Item Difficulty Levels.Journal of educational sciences: theory & practice (JESTP-DOI) 16(1), 319-330. 2016.
[17] G. Buck. Assessing Listening. Cambridge: Cambridge University Press 2001. In Soureshjan, K.H. Item Sequence on Test Performance: Easy Items First? Language Testing in Asia 3(1), 2-11. 2011.
[18] M. V. Santelices & M. Wilson. The Relationship between differential item functioning and item hardy: An issue of methods? Item response theory approach to differential item functioning. Educational and Psychological Measurement 72(1), 5–36. 2012.
[19] K.H. Soureshjani.Item Sequence on Test Performance: Easy Items First? Language Testing in Asia 3(1), 46-51.2011.
[20] K. MacNicol. Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript. Educational Testing Service, Princeton, New Jersey (1956). In S. N. N. Ollennu & Y. K. A. Etsey. Impact of Item Position in Multiple-choice Test on Student Performance at the Basic Education Certificate Examination (BECE) Level. Universal.Journal of Educational Research, 3(10): 718-723. 2015.
[21] L. A Shepard. The challenges of assessing young children appropriately 1997. In M.C. Katheleen. Educational Psychology. Shefield: Dubuque Inc 12th ed. 2012.

[22] M. O. Soyemi. Effect of item position on performance on multiple-choice tests. Unpublished M.Ed. dissertation, University of Jos . 1980.
[23] C. M. Anikweze. Measurement and evaluation for Teacher education, (2nd Ed.) Enugu, SNAAP Press 2016.
[24] Exams and record, department of Economics in Federal, State and Private Colleges of Education, Enugu State Enugu State, Nigeria, 2019.
[25] C. A. Ugodulunwa & U. P Okolo. Effects of formative assessment on mathematics Test Anxiety and Performance of Senior Secondary School Students in Jos, Nigeria. Journal of Research & Method in Education (IOSR-JRME) 2(5), 38-47. 2015.
[26] C. M. Anikweze. Measurement and Evaluation for Teacher Education, (2nd Ed.) Enugu, SNAAP Press 2016
[27] R. J. Zwick. A Review of ETS differential item functioning assessment procedures: Flagging principles, minimum sample size requirements, and criterion refinement (Research Report). Educational Testing Service, 2012.
[28] Ö. Çokluk, E. Gül, & C. Doğan-Gül. Examining Differential Item Functions of Different Item Ordered Test Forms According to Item Difficulty Levels.Journal of educational sciences: theory & practice (JESTP-DOI) 16(1), 319-330. 2016.
[29] M. V. Santelices & M. Wilson. The Relationship between differential item functioning and item hardy: An issue of methods? Item response theory approach to differential item functioning. Educational and Psychological Measurement 72(1), 5–36. 2012.
[30] O. I. Margaret & U. I. Victor. Effect of Test Item Arrangement On Performance In Mathematics Among Junior Secondary School Students In Obio/Akpor Local Government Area Of Rivers State Nigeria. British Journal of Education 8(5), 1-9. 2017.

Appendix A[1]

Table 3: DIF Results of Items in the Test Forms Easy-to-Hard and Hard-to-Easy Versions, based on CTT and IRT Methods

	CTT-Based Methods						IRT-Based Methods			
	MH	**DIF**	**TID**	**DIF**	**LR**	**DIF**	**Raju's Area**	**DIF**	**Chi-square**	**DIF**
Items		**Level**		**Level**		**Level**		**Level**		**Level**
1	0.882	A	0.221	C	0.892	A	**0.020**	C	**0.026**	B
2	0.704	A	0.844	A	0.159	A	**0.021**	C	**0.032**	B
3	0.673	A	0.606	A	0.504	A	**0.011**	B	**0.045**	B
4	**0.013**	A	0.616	A	0.530	A	**0.013**	C	**0.023**	C
5	0.431	B	0.669	A	0.865	A	**0.020**	C	**0.037**	B
6	0.642	A	0.651	A	0.729	A	**0.065**	A	0.191	A
7	**0.034**	**B**	**0.066**	**B**	**0.062**	**A**	**0.001**	**C**	**0.025**	**B**
8	0.686	A	0.642	A	0.961	A	**0.011**	C	**0.016**	C
9	0.535	B	0.748	A	0.331	C	0.334	A	0.646	A
10	0.654	A	0.073	A	**0.013**	B	**0.019**	C	**0.011**	C
11	0.457	B	0.643	A	0.503	A	**0.017**	C	0.120	B
12	0.465	B	0.806	A	0.495	A	0.854	A	0.692	A
13	**0.031**	**B**	**0.004**	**B**	**0.065**	**A**	**0.004**	**C**	**0.023**	**B**
14	0.547	B	0.699	A	0.363	C	0.115	C	**0.014**	C
15	0.643	A	**0.032**	C	0.974	A	**0.015**	B	**0.054**	A
16	0.452	B	0.660	A	0.839	A	**0.032**	B	0.138	A
17	0.311	C	0.742	A	0.495	A	0.161	A	0.263	A
18	0.491	B	0.464	B	0.132	B	0.155	A	0.225	A
19	0.911	A	0.574	B	**0.028**	B	0.372	A	0.119	A
20	0.823	A	0.699	A	0.638	A	**0.57**	B	0.158	A
21	0.462	B	0.621	A	0.892	A	0.110	C	**0.038**	B
22	0.693	A	0.544	B	0.059	A	0.601	C	**0.038**	B
23	0.390	C	0.606	A	0.504	A	**0.011**	B	**0.028**	B
24	0.261	C	0.716	A	0.530	A	0.150	C	0.136	C
25	0.642	A	0.669	A	0.865	A	**0.009**	C	**0.033**	B
26	**0.032**	**B**	**0.051**	**B**	**0.029**	**B**	**0.051**	**C**	**0.054**	**B**
27	0.361	C	0.666	A	0.462	A	**0.038**	B	0.165	A
28	0.512	B	0.742	A	0.961	A	**0.003**	C	0.111	C
29	0.543	B	0.848	A	0.331	A	0.334	A	0.544	A
30	0.804	A	0.773	A	**0.013**	C	**0.002**	C	**0.011**	C
31	0.731	A	0.643	A	0.503	A	**0.007**	C	0.126	B
32	0.561	B	0.606	A	0.495	A	0.852	A	0.678	A
33	0.311	C	0.604	A	0.495	A	0.444	A	0.186	A
34	0.512	B	0.699	A	**0.063**	A	**0.004**	C	0.084	C
35	0.531	B	0.632	A	0.974	A	**0.013**	B	0.159	A
36	0.261	C	0.060	A	0.839	A	**0.025**	B	0.136	A
37	0.433	B	0.442	A	0.495	A	0.151	A	0.265	A
38	0.721	A	0.864	B	0.232	B	0.135	A	0.222	A
39	0.483	B	0.774	B	**0.028**	B	0.373	A	0.119	A
40	**0.057**	**B**	**0.064**	**B**	**0.032**	**B**	**0.035**	**C**	**0.061**	**B**
No of items with DIF		**4**		**4**		**9**		**24**		**18**

Appendix A[2]

Table 4: DIF Results of Items in the Test Forms Easy-to-Hard and Random Versions, based on CTT and IRT methods

	CTT-Based Methods						IRT-Based Methods			
	MH	**DIF**	**TID**	**DIF**	**LR**	**DIF**	**Raju's Area**	**DIF**	**Chi-square**	**DIF**
Items		**Level**		**Level**		**Level**		**Level**		**Level**
1	0.882	A	0.221	C	0.892	A	**0.020**	C	**0.026**	B
2	0.704	A	0.844	A	0.159	A	0.221	C	**0.032**	B
3	0.673	A	0.606	A	0.504	A	**0.011**	B	**0.045**	B
4	**0.013**	A	0.616	A	0.530	A	**0.013**	C	0.123	C

5	0.431	B	0.669	A	0.865	A	0.120	C	**0.037**	B
6	0.642	A	0.651	A	0.729	A	**0.065**	A	0.191	A
7	**0.034**	**B**	**0.066**	**B**	**0.062**	**A**	**0.001**	**C**	0.225	C
8	0.686	A	0.642	A	0.961	A	0.111	C	**0.016**	C
9	0.535	B	0.748	A	0.331	C	0.334	A	0.646	A
10	0.654	A	0.073	A	**0.013**	B	**0.019**	C	**0.011**	C
11	**-0.007**	C	**-0.043**	A	0.503	A	**0.017**	C	0.120	B
12	0.465	B	0.806	A	0.495	A	0.854	A	0.692	A
13	**0.031**	**B**	**0.004**	**B**	**0.065**	**A**	**0.004**	**C**	**0.023**	**B**
14	0.547	B	0.699	A	0.363	C	0.115	C	**0.014**	C
15	0.643	A	**0.032**	C	0.974	A	**0.015**	B	**0.054**	A
16	0.452	B	0.660	A	0.839	A	**0.032**	B	0.138	A
17	0.311	C	0.742	A	0.495	A	0.161	A	0.263	A
18	0.491	B	0.464	B	0.132	B	0.155	A	0.225	A
19	0.911	A	0.574	B	**0.028**	B	0.372	A	0.119	A
20	0.823	A	0.699	A	0.638	A	**0.57**	B	0.158	A
21	0.462	B	0.621	A	0.892	A	0.110	C	**0.038**	B
22	0.693	A	0.544	B	0.059	A	0.601	C	**0.038**	B
23	0.390	C	0.606	A	0.504	A	**0.011**	B	**0.028**	B
24	0.261	C	0.716	A	0.530	A	0.150	C	0.136	C
25	0.642	A	0.669	A	0.865	A	**0.009**	C	0.233	A
26	**0.032**	**B**	**0.051**	**B**	**0.029**	**B**	**0.051**	**C**	**0.054**	**B**
27	0.361	C	0.666	A	0.462	A	**0.038**	B	0.165	A
28	0.512	B	0.742	A	0.961	A	**0.003**	C	0.111	C
29	0.543	B	0.848	A	0.331	A	0.334	A	0.544	A
30	0.804	A	0.773	A	**0.013**	C	**0.002**	C	**0.011**	C
31	0.731	A	0.643	A	0.503	A	**0.007**	C	0.126	B
32	0.561	B	**0.006**	A	0.495	A	0.852	A	0.678	A
33	0.311	C	0.604	A	0.495	A	0.444	A	0.186	A
34	0.512	B	0.699	A	**0.063**	A	**0.004**	C	0.084	C
35	0.531	B	0.632	A	0.974	A	0.213	B	0.159	A
36	0.261	C	0.060	A	0.839	A	**0.025**	B	0.136	A
37	0.433	B	0.442	A	0.495	A	0.151	A	0.265	A
38	0.721	A	0.864	B	0.232	B	0.135	A	0.222	A
39	**-0.003**	B	0.774	B	**0.028**	B	0.373	A	0.119	A
40	**0.057**	**B**	**0.064**	**B**	**0.032**	**B**	0.235	**A**	**0.061**	**B**
No of items with DIF		**6**		**7**		**9**		**20**		**15**

Appendix A[3]

Table 5: DIF Results of Items in the Test Forms Hard-to-Easy and Random Versions, based on CTT and IRT Methods

	CTT-Based Methods						IRT-Based Methods			
	MH	**DIF**	**TID**	**DIF**	**LR**	**DIF**	**Raju's Area**	**DIF**	**Chi-square**	**DIF**
Items		**Level**		**Level**		**Level**		**Level**		**Level**
1	0.882	A	0.221	C	0.892	A	0.720	A	0.226	B
2	0.704	A	0.844	A	0.159	A	0.521	B	0.232	B
3	0.673	A	0.606	A	0.504	A	**-0.011**	B	0.145	B
4	0.613	A	0.616	A	0.530	A	0.713	A	**0.023**	C
5	0.431	B	0.669	A	0.865	A	0.520	B	0.237	B
6	0.642	A	0.651	A	0.729	A	0.665	A	0.191	A
7	**0.034**	B	**0.066**	**B**	**0.062**	**A**	**-0.001**	**C**	0.225	**B**
8	0.686	A	0.642	A	0.961	A	**0.811**	A	0.216	C
9	0.535	B	0.748	A	0.331	C	0.334	A	0.646	A
10	0.654	A	**0.073**	A	**0.013**	B	**0.719**	A	0.211	C
11	0.457	B	0.643	A	0.503	A	**-0.017**	C	0.120	B
12	0.465	B	0.806	A	0.495	A	0.854	A	0.692	A
13	**0.031**	**B**	**0.764**	**A**	0.765	**A**	0.804	**A**	**0.023**	**B**
14	0.547	B	0.699	A	0.363	C	0.115	C	0.214	C
15	0.643	A	**0.032**	A	0.974	A	**-0.015**	B	0.254	A
16	0.452	B	0.660	A	0.839	A	0.532	B	0.138	A
17	0.011	A	0.742	A	0.495	A	0.161	A	0.263	A
18	0.491	B	0.464	B	0.132	B	0.155	A	0.225	A
19	0.911	A	**0.074**	B	**0.028**	B	0.372	A	0.119	A
20	0.823	A	0.699	A	0.638	A	0.657	B	**0.058**	A
21	0.462	B	0.621	A	0.892	A	0.110	C	0.238	B
22	0.693	A	0.544	B	0.059	A	0.601	C	**0**.138	B
23	0.590	B	0.606	A	0.504	A	0.511	B	0.228	B
24	0.861	A	0.716	A	0.530	A	0.150	C	0.136	C
25	0.642	A	0.669	A	0.865	A	**-0.009**	C	0.233	B
26	**0.032**	B	0.751	B	**0.029**	**B**	0.751	**C**	0.154	**B**
27	0.761	A	0.666	A	0.462	A	0.738	B	0.165	A
28	0.512	B	0.742	A	0.961	A	**-0.003**	C	0.111	C
29	0.543	B	0.848	A	0.331	A	0.834	A	0.544	A

30	0.804	A	0.773	A	**0.013**	C	0.672	A	0.211	C
31	0.731	A	0.643	A	0.503	A	**-0.047**	A	0.126	B
32	0.561	B	0.606	A	0.495	A	0.852	A	0.678	A
33	0.811	A	**0.004**	A	0.495	A	0.444	A	0.086	A
34	**0.012**	B	0.699	A	**0.063**	A	0.764	A	**0.084**	C
35	0.531	B	0.632	A	0.974	A	**-0.013**	B	0.159	A
36	0.561	B	0.060	A	0.839	A	0.525	B	0.136	A
37	0.833	A	0.442	A	0.995	A	0.651	A	0.265	A
38	0.721	A	0.864	A	0.832	B	0.935	A	0.222	A
39	0.583	B	0.774	A	**0.**528	B	0.173	A	**0.019**	A
40	0.557	**B**	0.564	**B**	0.532	**B**	0.235	**A**	0.661	**B**
No of items with DIF		**4**		**5**		**7**		**8**		**5**

Source: *Researchers Field Work*

 Salihu Abdullah **GALLE**

 **GBANDE**,Samson Samuel Kale

 **IBRAHIM,** Suleiman Agahu