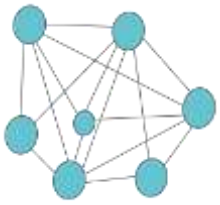
European Journal of
Educational & Social Sciences
Volume 3 Issue 1
May 2018
ISSN: 2564-6621



Article

Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results

Lydia Ijeoma Eleje [1] **Frederick Ekene Onah** [2]

Chidiebere Christopher Abanobi [3]

Manuscript information:

Received: April 24, 2018
Revised: May 25, 2018
Accepted: May 26, 2018

Author 1
Corresponding author, Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, NIGERIA
E-mail: ijeomaexcite@gmail.com

Author 2
Department of Life Science Education, Faculty of Education, Imo State University, Owerri NIGERIA
E-mail: fredonah2@yahoo.com

Author 3
Department of Educational Psychology, Federal College of Education (T), Asaba, NIGERIA
E-mail: abanobichidiebere@gmail.com

Abstract

In this study a comparison was made on DQEST item parameters and test statistics results estimated using Classical Test Theory (CTT) approach and Item Response Theory (IRT) three parameter logistic model (3PLM) to find out the similarities and differences in the two frameworks. 517 randomly selected senior secondary three (SS3) economics students comprised the sample. Three research questions guided the study. Responses obtained from SS3 economics students in 50 multiple choice items of Diagnostic Quantitative Economics Skill Test (DQEST) were used for the analysis. DQEST items certified the unidimensionality, local independence and model-data fit assumptions. Then results from CTT and IRT analyses were compared. In terms of very difficult item and item that discriminate poorly, CTT were found not to be comparable with the 3PLM the most appropriate model for DQEST data. The calculated reliability value for CTT was found to be low when compared to that generated by 3PLM. Therefore, it could be concluded that there was disparity between CTT approach and 3 parameter IRT model in terms of item parameters and test statistics. Thus IRT model with the best data fit should be employed for an enhanced test validity and reliability.

***Keywords*:** Item response theory, classical test theory, item parameters, local independence, unidimensionality.

Cite as:

Eleje, L. I.; Onah, F. E. & Abanobi, C. C. (2018). Comparative study of classical test theory and ıtem response theory using diagnostic quantitative economics skill test ıtem analysis results. *European Journal of Educational and Social Sciences, 3 (1),* 57 - 75.

INTRODUCTION

The essence of using tests and other evaluation instruments during the instructional process according to Kolawole (2010) is to guide, direct and monitor students' learning and progress towards attainment of course objectives. The goal of using tests in teaching and learning process will be accomplished only if such test is of good quality. Thus constructing valid and reliable tests is very important in assessing the students' performance. The quality of any test, for example diagnostic test, and the information the test generates is determined through item analysis of students' responses at any examinations. Item analysis according to Adedoyin and Mokobi (2013) is a process which examines students' responses to individual test items in order to assess the quality of those items and of the test as a whole. The two approaches commonly used for analysis of test items are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT has been the foundation for measurement theory for decades. The conceptual foundations, assumptions and extensions of the basic premises of CTT have allowed for development of some excellent psychometrically sound scales in the assessment practices of educational bodies in the world.

Despite the usefulness of classical test theory and models in psychometric methods, certain shortcomings underlying psychological testing and measurement procedures for test construction have been recognized and discussed. One of the shortcomings of CTT is that classical item statistics –item difficulty and item discrimination– depend on the particular examinee samples from which they were obtained (Cappelleri, Lundy & Hays, 2014). That is, test items look easy when administered to bright examinees, and harder when administered to less capable examinees. A consequence of this dependence on a specific sample of examinees is that these item statistics are only useful when constructing tests for examinee populations that are similar to the sample of examinees used in calibrating the test items. Unfortunately, one cannot always be sure that the population of examinees for whom a test is intended is similar to the sample of examinees used in obtaining the item statistics. Unlike CTT, item statistics that are invariant over examinee samples is one of the goals of modern test theory. The concept of invariance according to Ojerinde (2013) demands that the estimate of the parameters of an item across two or more group of populations of interest that are different in their abilities must be the same.

Another shortcomings of CTT according to Cappelleri, Lundy and Hays (2014), is that comparisons of examinees on the test score scale are limited to situations where examinees are administered the same (or parallel) tests. The seriousness of this shortcoming is clear when it is recognized that examinees often take different forms of a test. When several forms of a test that vary in difficulty are used, examinee scores across nonparallel forms are not comparable unless one makes use of equating procedures that are complex to implement in practice, especially with classical equating methods (Stage, 1998). As test scores are sample dependent, that is test scores depend on the set of items administered, they are not an adequate basis for score reporting or using norms tables.

Traditionally, the proficiency of individual examinees is reported in terms of number of items answered correctly. This constitute a limitation or weakness with CTT approach, in

that students with the same number of items answered correctly may have different response patterns (i.e., correct answers on different items) and, thus, may not have the same level of proficiency measured by the test. Also reports related to the quality of test items, are normally limited to indices of item difficulty (proportion of correct answers on the item) and item discrimination (Cappelleri, Lundy & Hays, 2014). But in CTT approach the problem with such indices is that they depend on the group of examinees being tested and, thus, do not adequately reflect the measurement quality of the test items. This problems that occur with CTT analysis of the examinees' proficiency and quality of test items are successfully addressed in the framework of Item Response Theory (IRT). For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information.

Hence, what is needed is an approach to ability estimation that is not test dependent and the influence of the particular items on the test administered to the examinee accounted for. It is necessary because an examinee may score high on an easy test or lower on a hard test, but there was a more fundamental ability that the examinee brings to any given testing situation that does not change as a function of the sample of items administered. This fundamental characteristic of the examinee is of interest to the psychologist and to IRT model.

The three different models of IRT are; one parameter logistic model (1PLM or Rasch model), two parameter logistic model (2PLM) and three parameter logistics model (3PLM). Taking these differences into consideration, the researchers in this study made a comparison between CTT and IRT logistic models in terms of item parameters. In an earlier study by Esomonu and Eleje (2017) Diagnostic Quantitative Economics Skill Test (DQEST) was developed and validated using IRT, and the test was found to be of good quality, valid and highly reliable. In this study a comparison was made on DQEST between item analysis results based on CTT and item analysis results based on the three IRT logistic models to find out the similarities and differences in the parameters estimated using these two frameworks. This paper therefore answer pertinent questions such as:

(a) To what extent are the IRT assumptions met for the DQEST data?
(b) How comparable are the CTT and IRT logistic models in terms of DQEST item parameters?
(c) How comparable are the CTT and IRT 3PLM test statistics for DQEST?

LITERATURE REVIEW

The following sub-headings guided the review of literature.

Classical Test Theory (CTT)

Classical test theory (CTT) is a theory about test scores that described how error can influence the observed scores or measurement. It introduces three concepts-test score (often called the observed score), true score (T), and error score (E). This is often expressed

mathematically in a simple linear model as X = T + E. Observed score 'X' is the simple sum of a True Score 'T' (which reflects the true amount of the attribute possessed by the person being measured at the time of measurement which is always contaminated by random errors) and an Error Score 'E' (which reflect the effect of extraneous influences of the measurement process at the time of measurement). According to Lord (1980), these random errors can result from several factors such as guessing, fatigue or stress. Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items. Therefore the equation is not solvable unless some simplifying assumptions are made.

Assumptions of Classical Test Theory

The three major assumptions that underlines CTT according to Lord (1980) are (a) true scores and error scores from the same test are uncorrelated, that is, they have a correlation of zero. Hence, the variance of the observed score is expected to be equal to the sum of the variances of the true and error scores. (b) The average error score in the population of examinees is zero. This means that these random errors over many repeated measurements are expected to cancel out in the long term run leaving the expected mean of measurement errors to be equal to zero. Once the error is zero, the observed score is equal to the true score. (c) Error scores on the parallel tests are uncorrelated. Lord (1980) went further to posit that in the definition of parallel tests in CTT, two tests of X and X^1 are considered parallel if the expected values of the two observed scores X and X^1 are equal (ie $E[X] = E[X^1]$) indicating that the two observed scores X and X^1 have the same true score $[T=T^1]$ and equal observed variances $\delta^2[X] = \delta^2[X^1]$.-The error variance for the two parallel scores are usually equal for every population of examines.

Reliability of a test in the CTT is then determined by the correlation coefficient between the observed scores on two parallel measurements. As the reliability of a measurement increases, the error variance becomes relatively smaller (Adedoyin, 2010; Ojerinde, 2013)). When the error variance is relatively small, an examinee's observed score is very close to the true score. However, when error variance is relatively large, observed score gives a poor estimate of the true scores (Lord, 1980).

Item Response Theory (IRT)

Item response theory (IRT) is a collection of measurement models that attempt to explain the connection between observed item responses on a scale and an underlying construct. Specifically, IRT models are mathematical equations describing the association between subjects' levels on a latent (hidden or dormant) variable and the probability of a particular response to an item, using a non-linear monotonic function (Cappelleri, Lundy & Hays, 2014; Hays, Bjorner, Revicki, Spritzer & Cella, 2009). IRT according to Ojerinde (2013) attempts to model the ability of an examinee's and the probability of answering a test item correctly based on the pattern of responses to the items that constitute a test. As in classical test theory, IRT requires that each item should be distinct from the others yet should be similar and consistent with them in reflecting all important respects of the underlying

attribute or construct. IRT makes it possible to scale test items for difficulty, to design parallel forms of tests, and to provide for adaptive computerized testing.

The purpose of IRT is to propose models that permit to link this latent trait to some observable characteristics of the examinee, especially his/her faculties to correctly answering to a set of questions that form a test (Bichi, Embong, Mamat & Maiwada, 2015; Magis 2007). IRT item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (asymptote) and they are estimated directly using logistic models instead of proportions. There are a number of IRT models varying in the number of parameters and whether they handle dichotomous only or polytomous items more generally (Cappelleri, Lundy & Hays, 2014). Three most commonly used IRT models are; one parameter logistic model (1PLM or Rasch model), two parameter logistic model (2PLM) and three parameter logistics model (3PLM). The characteristics of IRT models are summarized by Hambleton and Swaminathan (1985) as first, an IRT model must specify the relationship between the observed response and underlying unobservable construct. Secondly, the model must provide a way to estimate scores on the ability. Thirdly, the examinee's scores will be the basis for estimation of the underlying construct. Finally, an IRT model assumes that the performance of an examinee can be completely predicted or explained from one or more abilities.

Assumptions of IRT

Prior to estimating a latent trait model, it is important to evaluate its underlying assumptions. The two basic assumptions - unidimensionality and local independence are often evaluated (Cappelleri, Lundy & Hays, 2014). A test data can only be useful for an IRT model estimation only if these assumptions are met.

1 Unidimensionality: The assumption of unidimensionality assumes that a set of items and/or a test measure(s) only one latent trait (Kyung, 2013). This implies that the performance of each examinee is assumed to be governed by a single factor, referred to as ability. Since individuals' cognitive and personal characteristics, influence test performance and cannot often be controlled; it is not always possible to meet this assumption. One can then talk about the unidimensionality of a test only when there is just one dominant ability in it (Hambleton, Swaminathan & Rogers, 1991).

To satisfy this assumption, one can apply any these eleven methods for testing for unidimensionality as cited by Ojerinde and Ifewulu (2012); Cronbach analysis test, Factor analysis, Eigenvalue test, Random baseline test, Biserial test, Factor loading test, Congurence test, Part/Whole test, Communality test, Vector frequency test and Confirmatory factor analysis (F.A) and Structural equation modelling (SEM) test, using the SPSS package. A support for the unidimensionality of the items in the scale is provided when the model fits the data well and there are no noteworthy residual correlations (i.e., no such correlations greater than or equal to 0.20) (Ojerinde, 2013). Any violation of this assumption would result in inadequacy of the model in describing the data and hence unreliable estimation of the examinee's ability. Therefore, the correct specification of the number of the latent dimensions

is directly tied to the construct validity of the test (Rijn, Sinharay, Haberman & Johnson, 2016). In this study, DQEST is assessed for unidimensionality.

2 Item Local Independence: Local independence refers to the assumption that there is no statistical relationship between examinees' responses to the pairs of items in a test, once the primary trait measured by the test is removed (Kyung, 2013). It implies that the probability of an examinee getting an item correct is unaffected by the answer given to other items in the test. Local independence according to Ojerinde (2013) does not mean that items do not correlate with each other, but that performance on different items is independent but conditional on the student's ability. Thus, the probability that a student will answer correctly any two items must be the product of the probability that the student will answer correctly each separate item. Also, the association between two items should not differ significantly from zero, otherwise, it may be said that the responses to the items are influenced by other extraneous factors other than what the instrument is designed to measure (Ojerinde, 2013).

Item Characteristic Curve of IRT

An item characteristic curve (ICC) is a mathematical function that relates to the probability of endorsing an item (for a dichtomous response) or responding to a particular category of an item (for a polytomous response) for individuals with a given level of the attribute. This probability is independent on the distribution of examinees of interest. Since the probability is independent of how many other examinees are located at the same point on the ability in the examinees population. The various IRT models, which are variations of logistic (i.e., non-linear) models, are simply different mathematical functions for describing ICCs as the relationship of a person's level on the attribute and an item's characteristics (e.g., difficulty, discrimination) with the probability of a specific response on that item measuring the same attribute (Ani, 2014; Cappelleri, Lundy & Hays, 2014).

When ICCs are plotted the ability of the examinee is denoted by theta (θ) on the *x*-axis, while the probability of an examinee correctly answering the question is denoted by P(θ) on the y-axis. The ICC is monotonic and takes the shape of an S – shaped curve that is normal ogive (θ). It has three parts, the lower asymptote, the upper asymptote and the middle or rapidly rising part of the ICC. The number of parameters required to determine an item characteristic curve to Ojerinde (2013) depends on the particular logistic model.

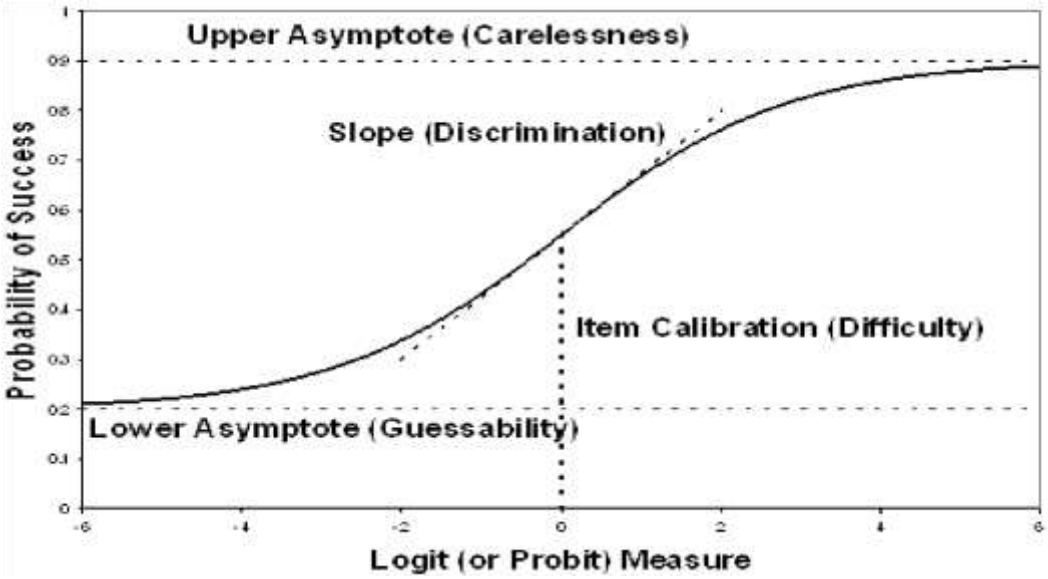


Fig 1. Example of Item Characteristics Curve (ICC)

Analysis of Model-Data fit

The analysis of model-data fit is a check on internal validity. Within the latent trait test model, the internal validity of a test is assessed in terms of the model with the most compatibility to the data. If the data is compatible to the model, then the item is valid (Kyung, 2013). The IRT has three models: one-parameter, two-parameter and three parameter models. Data fit to the model, implies that item discriminations are uniform and substantial, that there are no errors in item scoring. -2Log likelihood value is commonly used to check the model-data fit. Comparing the values from different models can indicate which model represents a better fit. However, the smallest -2Log likelihood value is the best. (Thorpe & Favia, 2012).

Comparison of CTT and IRT

IRT provides a richer set of tools for test developers. It provides a psuedo-guessing parameter that has no common analog in CTT it also provides a means to assess degree of measurement equivalence at various points on the score scale and based on different sets of items. The item analysis statistics provided by both IRT and CTT are fairly comparable, but IRT provides an additional item characteristic curve and a more sophisticated mechanism for conceptualizing measurement error.

While IRT does everything that CTT does and more, there are a few possible advantages of CTT. CTT is simple. Whereas IRT requires relatively obscure software, CTT item analysis is easy to conduct in common statistical packages. Also CTT statistics are easily computed, while IRT statistics must be estimated, which is more complex in its own right and requires a thorough model-data-fit analysis. Moreover, explaining test properties to

a lay man is considerably simple using CTT. The other advantage of CTT is that it has few assumptions. CTT may be a better choice in situations where IRT models do not fit well because of violations of the assumptions or shape of the model (Mead & Meade, 2010).

Bichi, Embong, Mamat and Maiwada (2015) gave the comparison of CTT and IRT as shown below (Table 1):

Table 1: Comparison of CTT and IRT

Area	**CTT**	**IRT**
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e easy to meet with test data)	Strong (i.e more difficult to meet with data)
Item-ability relationship	Not specified	Item characteristics functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformation test score scale)	Ability scores are reported on scale $-\infty$ to $+\infty$ (or transformed scale)
Invariance of item and person statistics	Item and person parameters are sample dependent	Item and person parameters are sample independent, if model fits test data
Item statistics	*p, r*	*b, a* and *c* (for the three-parameters model) plus corresponding item information functions
Sample size (for item parameters estimation)	200 to 500 (in general)	Depends on the IRT model but larger samples. That is over 500, in general are needed

From the table above, one can then say that, IRT seems to be superior to CTT in many ways. It is conceptually superior, IRT provides a richer selection of tools for test developers, and IRT has advantages that are hard to quantify, like greater flexibility, invariance of the parameters, and providing adequate statistical models. The only area where the superiority of IRT is not obvious is for smaller samples or tests which were not unidimensional or where, for some reasons, the data do not fit the IRT model, and perhaps in the area of ease of use (Mead & Meade, 2010).

Researchers like, Stage (1998), Mead and Meade (2010), Ojerinde (2013) and Guler, Uyanik and Teker (2014) have compared CTT and IRT framework item analysis results. Stage (1998) compared the item statistics from the CTT framework with those from the IRT framework and examined the stability from pretest to regular test of the two sets of item statistics. The overall conclusion from the study is that the prediction from pretest data to regular test data is very good but that is true for CTT as well as for IRT.

Mead and Meade (2010) as well compared test construction using CTT and IRT in several sample sizes (from N=20 to N=5000) and degrees of representativeness (represented by selecting the top 20%, 40%, 60%, 80% or 100% of a population) using a Monte-Carlo simulation design. They also concluded that test construction using either CTT or IRT produces empirically similar exams and IRT is only preferred when there is a target test information function.

Ojerinde (2013) conducted a study to evaluate the psychometric utility of data obtained using the two models in the analysis of UTME Physics Pre-test so as to examine the results obtained and determine how well the two can predict actual test results and the degree of their comparability. It was found out that the indices obtained from both approaches gave valuable information with comparable and almost interchangeable results in some cases. The paper recommended that both IRT and CTT parameters should be used together in empirical determination of the validity of MCQ items to ensure a common basis of test item analysis in which the defect of one is compensated for by the other.

The above studies concluded that that CTT and IRT produce similar results. However a study by Guler, Uyanik and Teker (2014) identified an area of difference between CTT and IRT item analysis results. Guler, Uyanik and Teker (2014) in their study examined the similarities and differences in the parameters estimated using these two frameworks. They found that the highest correlations were available between CTT and 1parameter IRT model (0.99) in terms of item difficulty parameters, and between CTT and 2-parameter IRT model (0.96) in terms of item discrimination parameters. Although they identified 3-parameter model as the best model in terms of model-data fit, the lowest level of correlation was found between the 3-parameter model and CTT. In the light of these their findings, it may be said that there is not much difference between using 1 or 2-parameter IRT model and CTT. However, there is a significant difference between 3-parameter model and CTT. Does this apply to DQEST?

Thus, in this study a comparison was made on DQEST between item analysis results based on CTT and item analysis results based on IRT to find out the similarities and differences in the parameters estimated using these two frameworks.

METHOD

The study population was 917 senior secondary three (SS3) economics students in the 49 public secondary schools in Nigeria for 2016/2017 academic year, out of which 517 students were randomly selected as the study sample. The data for the study involve the responses given by the 517 SS3 economics students on fifty (50) multiple choice items of Diagnostic Quantitative Economics Skill Test (DQEST). That is, the data needed for the necessary CTT and IRT analysis. In order to satisfy the conditionality for the implementation of IRT, the fifty (50) multiple choice items of DQEST were first assessed to determine the dominance of the first factor (unidimensionality), secondly for local independence and thirdly for model-data fit. Also the results gotten from CTT and IRT DQEST item analysis were compared.

The method used to test for unidimensionality of the DQEST items was factor analysis. The factor analysis was done using SPSS. The eigenvalues and scree plot obtained from the factor analysis was investigated in order to check whether there was a dominant factor. The local independence of DQEST on the other hand, was assessed by conducting a tetrachoric correlation using Lisrel software. The result of tetrachoric correlation was checked to determine the level of compliance of the DQEST items with the assumption of local item independence. To check for model-data fit or the compatibility of the data with 1, 2 and 3 parameter logistic models, the -2 Log Likelihood values for the parameter logistic models were found. According to Thorpe and Favia (2012), the model with the smallest -2 Log Likelihood value has the best data fit. Then item parameters for the CTT and IRT analysis were derived with the Bilog MG software.

FINDINGS

Research question 1:

To what extent are the IRT assumptions met for the DQEST data?

Confirming the assumption of unidimensionality

1. Factor analysis

Table 2 shows that the first three eigenvalues for the 50 test items on the Diagnostic Quantitative Economics Skill Test (DQEST) were 6.623, 1.976, and 1.759. Since the first eigen value is substantially greater than the next, it appeared reasonable to conclude that the unidimensionality assumption for the IRT models were sufficiently met/satisfied for the DQEST data used in the study.

The questionnaire enabled us to understand which interactive strategies are used by instructors in search for finding the most effective ways for the students to express their knowledge and understanding based on their abilities, needs, and interests, and specific learning style of each learner. The results of the students responses to the question regarding the preferred forms of expression are as follows:

Table 2: Total Variance Explained by the result of factor analysis

Component	Initial Eigenvalues			Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %		Total	% of Variance	Cumulative %
1	6.623	13.246	13.246	26	.796	1.593	73.630
2	1.976	3.951	17.197	27	.786	1.571	75.201
3	1.759	3.518	20.716	28	.756	1.511	76.712
4	1.679	3.359	24.075	29	.729	1.458	78.170
5	1.615	3.230	27.304	30	.713	1.425	79.595
6	1.589	3.178	30.482	31	.690	1.380	80.975
7	1.511	3.023	33.505	32	.670	1.340	82.315
8	1.420	2.840	36.345	33	.659	1.319	83.634
9	1.372	2.743	39.088	34	.622	1.244	84.877
10	1.289	2.579	41.667	35	.589	1.177	86.055
11	1.270	2.541	44.208	36	.584	1.169	87.224
12	1.213	2.426	46.634	37	.563	1.126	88.349
13	1.154	2.308	48.942	38	.555	1.110	89.459
14	1.118	2.236	51.178	39	.538	1.076	90.535
15	1.082	2.163	53.341	40	.515	1.029	91.565
16	1.055	2.111	55.452	41	.496	.991	92.556
17	1.029	2.059	57.511	42	.490	.981	93.537
18	1.007	2.014	59.525	43	.476	.951	94.488
19	.966	1.932	61.457	44	.463	.926	95.413
20	.926	1.852	63.309	45	.446	.891	96.305
21	.911	1.823	65.131	46	.409	.818	97.123
22	.888	1.775	66.907	47	.406	.812	97.935
23	.887	1.773	68.680	48	.360	.720	98.655
24	.857	1.715	70.395	49	.344	.689	99.344
25	.821	1.642	72.037	50	.328	.656	100.000

Extraction Method: Principal Component Analysis.

2. Eigen value test

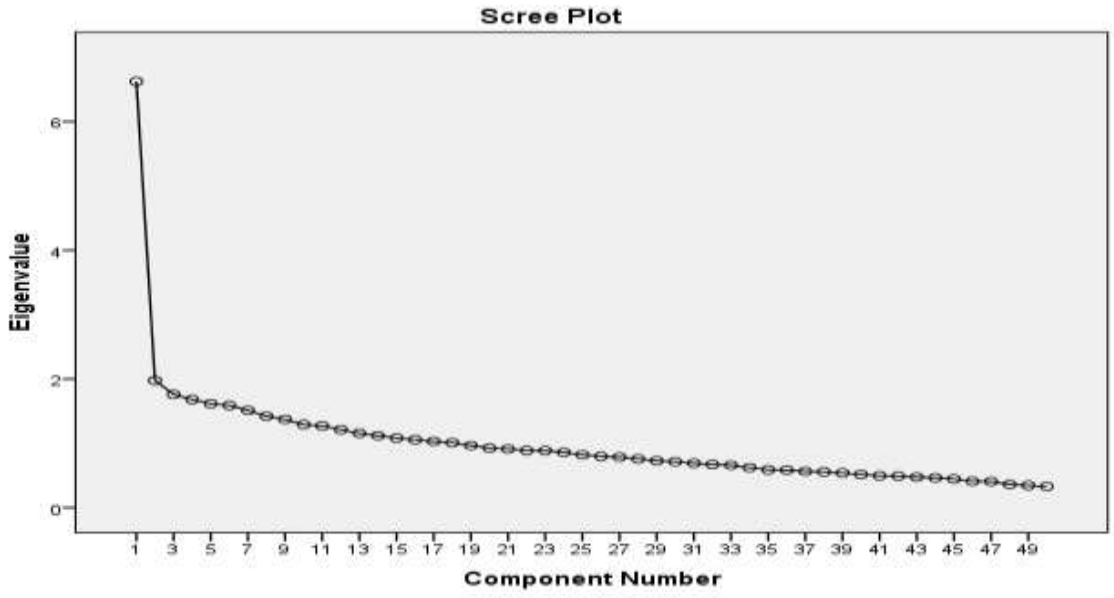


Fig 2: Scree plot of DQEST

Fig 2 shows the scree plot produced from the result of the eigen value test. As seen in Fig 1, the first eigen value was larger compared to the second factor, while the eigen value of the remaining factors are all about the same.

Confirming the assumption of local independence

1. Tetrachoric corelation

Table 3. Summary of tetrachoric correlations for DQEST

	item1	item2	item3	item4	item5	item6
item1	1.000					
item2	0.590	1.000				
item3	0.425	0.433	1.000			
item4	0.403	0.292	0.397	1.000		
item5	0.375	0.353	0.562	0.529	1.000	
item6	0.064	-0.007	0.003	-0.118	-0.155	1.000
item7	0.227	0.080	0.260	0.177	0.116	0.024
item8	0.129	0.110	0.247	0.392	0.434	0.045
item9	0.321	0.366	0.421	0.300	0.358	0.053
item10	0.323	0.215	0.438	0.301	0.405	-0.133
item11	0.266	0.102	0.262	0.272	0.225	0.107
item12	0.197	0.138	0.170	0.265	0.183	0.328

item13	-0.001	-0.100	0.092	0.230	0.314	-0.022
item14	0.264	0.285	0.306	0.291	0.377	-0.035
item15	-0.100	-0.035	-0.006	-0.142	-0.039	0.110
item16	0.334	0.329	0.336	0.266	0.367	0.051
item17	0.408	0.294	0.293	0.488	0.528	-0.165
item48	0.026	0.039	-0.096	0.095	-0.009	-0.158
item49	0.127	-0.014	0.243	0.208	0.086	0.088
item50	0.171	0.331	0.325	0.117	0.119	0.102

Table 4: Frequency Distributions of Tetrachoric Correlations for DQEST

Subject	**Correlation Coefficient**	**Frequency**	**Percentage**	**Remark**
Diagnostic Quantitative Economics Skill Test (DQEST)	Greater than 0 .500	14	1.15	
	0.450 - 0.500	23	1.89	
	0.300 – 0.449	192	15.75	
	0.200 – 0.299	277	22.72	Close to zero
	0.0 – 0.1	713	58.49	Very close to zero

Table 3 presents the summary of DQEST tetrachoric correlations. Also Table 4 shows the frequency distribution and percentage of DQEST tetrachoric correlations. As seen in Table 4, the percentage of correlation coefficients of DQEST items that are close to zero is 81.21%. This implies that greater number of DQEST items correlation coefficients are close to zero. Thus, diagnostic quantitative economics skill test items are locally independent.

Analysis of model-data fit

Table 5. The -2 Log Likelihood values of 1, 2, and 3 parameter logistic models

Model	**-2 Log Likelihood Values**
1PLM	30901.938
2PLM	30332.641
3PLM	30314.198

As observed in Table 5, the model with the lowest -2 Log Likelihood is three parameter logistic model (3PLM). Therefore, the 3PLM had the best model-data fit than the

1PLM and 2PLM. Thus, three parameter logistic (3PL) model is the best model for DQEST item and thus was used to in this study to estimate DQEST test statistics. The essence of using the best model fit for the items according to Thorpe and Favia (2012) is to ensure the validity of the test items.

Research question 2:

How comparable are the CTT and IRT logistic models in terms of DQEST item parameters?

The obtained difficulty parameter (b) for the 1PL model, the difficulty (b) and discrimination (a) parameters for the 2PL and 3PL IRT models, and difficulty (p) and discrimination (r) parameters for CTT are presented in Table 6.

Analysis in Table 6 revealed that the fourth item with the values of -2.18 and 81.8 for the 1 PLM and CTT respectively is the relatively easiest item. None of the items from 2PL and 3PL were found to be very easy. Item 15 of which the values are 2.28, 5.92 and 17.2 respectively for 1 PLM, 2PLM and CTT is the very difficult item, while item 48 the value of which is 3.80 is the very difficult for 3PLM. Table 6 also have the item discrimination values for the 1 PLM, 2PLM, 3PLM and CTT. The item that discriminate poorly for 2PLM and CTT, is item 15 of which the values are 0.15 and -0.03 respectively. On the other hand, item 39 with the value of 0.20 discriminate poorly for 3PLM. One parameter IRT model only provides estimates for item parameter of difficulty since it assumes fixed item discrimination.

Table 6. Item statistics (parameters) calculated by IRT and CTT

	IPLM	**2PLM**		**3PLM**		**CTT**	
Item	**b**	**a**	**b**	**a**	**b**	**p**	**r**
1	-1.80	0.61	-1.44	0.78	-0.76	77.4	0.29
2	-1.49	0.53	-1.32	0.58	-0.83	73.3	0.27
3	-1.27	0.86	-0.81	0.95	-0.51	69.6	0.41
4	-2.18	0.78	-1.48	0.85	-1.19	81.8	0.31
5	-1.77	0.87	-1.13	1.04	-0.82	77.0	0.36
6	1.29	0.18	2.88	1.59	2.05	29.2	0.03
7	1.92	0.35	2.36	1.21	1.94	21.1	0.21
8	-0.83	0.67	-0.65	0.81	-0.19	63.6	0.37
9	-0.47	0.62	-0.39	0.71	-0.01	57.6	0.36
10	-1.65	0.84	-1.08	0.93	-0.70	75.4	0.37
11	-0.24	0.62	-0.21	1.28	0.55	53.8	0.38
12	-0.44	0.69	-0.35	1.43	0.42	57.3	0.40
13	0.88	0.41	0.94	0.71	1.41	35.4	0.28

14	-0.42	0.62	-0.35	0.88	0.26	56.9	0.39
15	2.28	0.15	5.92	1.27	2.35	17.2	-0.01
16	0.81	0.82	0.50	1.20	0.73	36.4	0.47
17	-0.24	0.99	-0.19	1.40	0.13	53.8	0.50
18	0.70	0.73	0.47	1.94	0.78	38.1	0.44
19	0.89	0.38	1.02	0.64	1.40	35.2	0.24
20	-0.14	0.52	-0.14	0.69	0.36	52.2	0.33
21	-0.19	0.78	-0.16	1.11	0.26	53.0	0.44
22	0.20	0.63	0.13	0.81	0.46	46.4	0.37
23	1.01	0.25	1.72	0.95	2.14	33.3	0.14
24	2.02	0.21	3.97	1.60	2.46	19.9	0.07
25	0.19	0.97	0.07	1.28	0.35	46.6	0.51
26	0.23	0.58	0.17	0.73	0.51	45.8	0.37
27	0.26	0.50	0.22	1.06	0.88	45.5	0.33
28	0.29	0.73	0.17	1.20	0.59	44.9	0.45
29	0.67	0.47	0.64	1.10	1.17	38.7	0.32
30	0.62	0.70	0.42	1.63	0.86	39.5	0.45
31	-0.09	0.59	-0.09	1.45	0.69	51.3	0.37
32	1.32	0.25	2.16	1.08	2.06	28.8	0.14
33	0.73	0.20	1.47	0.77	2.61	37.7	0.10
34	1.43	0.42	1.51	1.36	1.57	27.3	0.28
35	0.74	0.50	0.67	2.22	1.17	37.5	0.34
36	1.61	0.28	2.38	0.54	2.63	25.0	0.16
37	0.81	0.53	0.70	1.36	1.05	36.4	0.36
38	0.20	0.55	0.15	0.74	0.57	46.4	0.37
39	2.01	0.28	3.02	0.20	2.51	20.1	0.14
40	1.40	0.19	2.97	1.29	2.66	27.7	0.06
41	1.92	0.34	2.40	1.21	1.20	21.1	0.19
42	0.18	0.40	0.19	0.60	0.87	46.8	0.26
43	0.07	0.47	0.06	0.56	0.43	48.5	0.32
44	0.07	0.37	0.09	0.68	1.11	48.5	0.25
45	0.33	0.56	0.29	0.80	0.80	44.3	0.33
46	0.35	0.67	0.23	1.60	0.77	43.9	0.42
47	0.69	0.29	1.00	1.15	1.73	38.3	0.20
48	1.04	0.16	2.60	0.56	3.80	32.9	0.02
49	1.21	0.39	1.35	0.68	1.62	30.4	0.26
50	0.75	0.60	0.59	0.83	0.86	37.3	0.42

Research question 3:

How comparable are the CTT and IRT 3PLM test statistics for DQEST?

Table 7. Comparison of IRT (3PLM) and CTT Test Statistics for the DQEST

	Reliability	SD	No of Items rejected because of a = <0.3 or > 2 for IRT or a = <0.3 or > 0.7 for CTT	No of items rejected because b = < -2 or > 2 for IRT or b = < 0.2 for CTT
IRT	0.8659	0.9350	2	8
CTT	0.6598	7.6830	11	15

Tables 7 revealed the test statistics for DQEST derived from the Item Response Theory (IRT) model and the Classical Test Theory (CTT) approach. The table indicated a great improvement in the DQEST statistics using the 3PL IRT model compared to the CTT approach. Total number of items rejected on the basis of the 'a' (discrimination) index was only 2 for the IRT model while 11 were rejected using CTT approach. The number of items rejected as a result of item difficulty 'b' was 8 for IRT model while 15 was rejected for using CTT approach.

As also seen in table 7, the reliability coefficients derived using Kuder-Richardson 21 formula for the IRT and CTT vary. The reliability coefficient as computed by Bilog MG in 3PL IRT model was given as 0.8659, while that of CTT was calculated as 0.6598. The calculated reliability value for CTT was low when compared to that generated by IRT model. This according to Ojerinde (2013) was as a result of the rejection of more items by CTT than the IRT. Ojerinde also pointed out that a negative discrimination value (r) would have accounted for the low reliability value encountered by CTT. Negative discrimination value occurs when low ability students performed better on an item than high ability students. Such items discriminated but in the negative (wrong) direction. Such items in test development and evaluation, is reviewed then replaced or amended for improvement of the test.

DISCUSSION

The overall results of this study has revealed that the IRT method used in analyzing DQEST data was more valid and reliable than the CTT approach. The findings of this study was in line with that of Ojerinde (2013) who conducted a study on the evaluation of the comparability of IRT and CTT item analysis results using UTME Physics Pre-test. As in Ojerinde's study, many DQEST items were rejected by the CTT approach than with IRT

model, with their difficulty and discrimination indices below the bench mark. The test statistics from both studies also indicated a good test statistics of IRT over CTT.

The comparability check on CTT and IRT parameters indicated that the easiest item from CTT were very comparable with especially the 1PL model. None of the items from 2PL and 3PL were found to be very easy. Compared with very difficult item, CTT were not comparable with the model that had the best DQEST data fit, which is the three parameter logistic model (3PLM). Also the item that discriminate poorly for CTT was similar to 2PL model, but not comparable with those from 3PL model. This was similar with the findings of Guler, Uynik and Teker (2014), who in their study pointed out that the IRT model that are not comparable with CTT approach was found to be 3PL model- the most appropriate model in terms of data-model fit.

Where IRT is used in carrying out item analysis of an instrument, assumptions of IRT model need to be satisfied or met. Conforming to IRT conditionality, according to Ojerinde (2013) will ensure the full employment of the principles of IRT in solving measurement problems. Local independence assessment shows that DQEST items were to a great extent locally independent. A significant number (81.21%) of DQEST items correlations were approximately zero. It means that the probability of a student getting DQEST item correct is unaffected by the answer given to other items in the test. This according to Ojerinde (2013) does not mean that items do not correlate with each other, but that performance on different items is independent but conditional on the student's ability. That is, items are not related and may not have acted as a clue to one another during the testing session.

The dimensionality assessment also shows that DQEST had one dominant factor. Such according to Rijn, Sinharay, Haberman and Johnson (2016) is an evidence that construct validity of the test is ensured. The smallest -2Log likelihood value of 3PLM indicates that the model represents the best model for DQEST data. If the data is compatible to the model, then the item is valid (Kyung, 2013). The assumption of local independence, unidimensionality and model data fit, were all fulfilled in this study and thus enhanced DQEST validity and reliability.

CONCLUSION

As noted in the findings of this study the calculated reliability for CTT was 0.66 and the empirical reliability of the overall test for IRT was 0.87. This according to Ceniza and Cereno (2012) and (Cherry, 2009) implied that DQEST reliability was high for IRT model since there was 87% certainty of the consistency of the test items in yielding approximately same result repeatedly. When compared in terms of test statistics and item parameters, the researchers can say that there was difference between 3PL model and CTT. However, there was a slight difference between using 2PLM and CTT. Thus IRT model with the best data fit should be employed for an enhanced test reliability. Therefore, it could be concluded that

there was disparity between CTT approach and 3 parameter IRT model in terms of item parameters and test statistics.

RECOMMENDATIONS

The researchers therefore recommended that examination bodies should use IRT model in test development process. Effort should be taken to ensure conformity of the conditions or assumptions of IRT model. The need for educational institutes to create more interest and awareness for students, teachers and stakeholders in the application of IRT was also recommended. This can be enhanced by including IRT as part of undergraduate curriculum as well as more local and international workshops and seminars for both post graduate students and academic staff.

REFERENCES

Adedoyin, C. (2010). Investigating the invariance of persons' parameter estimates based on classical test and item response theories. *An International Journal on Education Science*, 2(2), 107-113. Retrieved from http:// *krepublishers.com/...Journals/...2...2...Adedoyin.../IJES-2-2-107-10-033-Adedoyin-O...*

Adedoyin, O.O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4), 992-1011. Retrieved from http://www.aessweb.com/journal etail.php?id=5007

Ani, E.N. (2014). *Application of item response theory in the development and validation of multiple choice test in economics.* (Master's thesis). University of Nigeria, Nsukka.

Bichi, A.A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: a review of empirical studies. *Australian Journal of Basic and Applied Sciences*, 9(7), 549-556. Doi:10.13140/RG.2.1.1561.

Cappelleri, J.C., Lundy, J.J., & Hays, R. D. (2014). *Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures.* Doi: 10.1016/j.clinthera.2014.04.006

Esomonu, N.P.M. & Eleje, L.I. (2017). Diagnostic quantitative economics skill test for secondary schools: Development and validation using item response theory. *Journal of Education and Practice*, 8(22), 110-125. Retrieved from www.iiste.org

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/0013164498058003001

Guler, N., Uynik, G.K., & Teker, G.T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, 2(1), 1-6. Retrieved from http://iassr.org/journal

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory. Principles and application.* Retrieved from www.springer.com/gp/book/9780898380651

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA:Sage Publications

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research,* (18), 873–80.

Kolawole, E.B. (2010). *Principles of test construction and administration (Revised Edition).* Lagos: Bolabay Publications.

Kyung, T. H. (2013). *Windows software that generates IRT parameters and item responses: research and evaluation program methods (REMP)*. University of Massachusetts Amherst. Retrieved from https://www.umass.edu/remp/software/simcata/wingen/homeF.html

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Magis, D. (2007). *Influence, information and item response theory in discrete data analysis.* Retrieved from http://bictel.ulg.ac.be/ETDdb/collection/available/ULgetd-06122007-100147/.

Mead, A.D., & Meade, A.W. (2010). *Item selection using CTT and IRT with unrepresentative samples.* Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA. Retrieved from https://www.researchgate.net/...Classical_Test_Theory...Item.../Comparison-of-Classica...

Ojerinde, D. (2013). *Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results.* Retrieved from https://ui.edu.ng/sites/.../PROF%20OJERINDE'S%20LECTURE%20(Autosaved).pdf

Ojerinde, D., & Ifewulu, B. C. (2012). *Item unidimensionality using 2010 unified tertiary matriculation examination mathematics pre-test. A paper presented at the 2012 international conference of IAEA, Kazastan.* Retrieved from https://ui.edu.ng/sites/.../PROF%20OJERINDE'S%20LECTURE%20(Autosaved).pdf

Rijn, R.W.V., Sinharay, S., Haberman, S.J., & Johnson, M.S. (2016). *Assessment of fit of item response theory models used in large-scale educational survey assessments.* DOI: 10.1186/s40536-016-0025-3

Stage, C. (1998). *A comparison between item analysis based on item response theory and classical test theory.* A study of the SweSAT subtest ERC. (Educational Measurement). Umeå University, Department of Educational Measurement. Retrieved from www.edusci.umu.se/digitalAssets/60/60608_enr3098sec.pdf

Thorpe, G. L., & Favia, A. (2012). *Data analysis using item response theory methodology: an introduction to selected programs and applications.* Retrieved from http://digitalcommons.lidrary.umaine.edu/psy_facpub/20