Classical Test Theory (CTT) VS Item Response Theory (IRT):

AN EVALUATION OF THE COMPARABILITY OF ITEM ANALYSIS RESULTS

BY

Prof. 'Dibu Ojerinde, OON

Joint Admissions and Matriculation Board (JAMB), Abuja, Nigeria

Lecture Presentation at the Institute of Education

University of Ibadan

May 23, 2013

Classical Test Theory VS Item Response Theory: An Evaluation of the Comparability of Item Analysis Results

Abstract

Classical Test Theory (CTT) has served measurement practitioners for several decades as the foundation measurement theory. The conceptual groundwork, assumptions, and extension of the basic principles of CTT have allowed for the development of some excellent psychometrically sound scales. IRT is relatively new on stage and the advantage of item and test data invariance makes it to be preferable. Is it proper then to jettison CTT for IRT? The aim of this paper is 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. The paper also verified the conditions to be fulfilled for IRT to be usefully applied with real test data. Findings showed that the result obtained using the IRT model was found to be more suitable in multiple-choice questions (MCQs) in ability test but involved more complex mathematical estimation procedure than the classical approach. In the overall, 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.

Introduction

Classical Test Theory (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 the development of some excellent psychometrically sound scales in the assessment practices of educational bodies in Africa. This is owing to the simplicity of interpretation which can usefully be applied to examinees achievement and aptitude test performance (Hambleton, 1989). In the past 30 years or more, the field of educational measurement all over the world has undergone changes to meet increasing demand for valid interpretation of individual score from educational tests or examinations (Adedoyin 2010).

Despite the popularity of Classical Item Statistics as an integral part of standardized testing and measurement technology, it is fraught with many shortcomings. Among these are the values of standard item parameters that are invariant across groups of examinees that differ in ability. The concept of invariance 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. The quest for invariant estimate of ability has therefore led to the birth of Item Response Theory (IRT). IRT models are often referred to as latent trait models. The term latent is used to emphasize that discrete item responses are taken to be observable manifestations of hypothetical traits, constructs or attribute not directly observed, but that are manifest responses. IRT is generally regarded as an improvement over CTT. There is nothing CTT can accomplish that IRT cannot. The converse is however not true. For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information. Some applications such as computer adaptive testing are enabled more easily by IRT and may not be performed with CTT. The sophisticated information IRT provides has therefore given opportunity to the researchers to improve the reliability of assessment (Wikipedia, 2011).

Statement of the Problem

CTT has served practitioners for a long time with inherent attributes which have contributed immensely to the development of instrument for comparable evaluation of achievement. The property of invariance of person and item characteristics is very critical for objective measurement. Are these demands sufficient enough to jettison CTT for IRT? This should not be the case owing to the fact that CTT has worked for years in making testing decisions and generalizations. The conditions for implementation of IRT have to, as a first step be fulfilled before it can be effectively married into measurement scales for proper use and interpretation. This paper therefore prepares ground to answer pertinent questions such as:

(a) What are the conditionality that need to be fulfilled before IRT can be usefully applied?
(b) How comparable are the item analysis results of CTT and IRT?
(c) Can CTT psychometric data be complementary to that of IRT or vice versa?

Classical Test Theory (CTT)

CTT describes how error can influence observed scores or measurement. CTT is based on the true score theory which introduces three concepts – Test scores (often called Observed scores), true score (T) and error score E. This is often expressed mathematically as $X = T + E$

The true score, T reflects whether the examinee's amount of knowledge or ability is the true measurement of the examinee which is always contaminated by random errors. According to Lord (1980), these random errors can result from several factors such as guessing, fatigue or stress. The observed score is often called a fallible score because of the error contaminant. The true score is the score that would have been obtained if there were no error in measurement.

Assumptions of Classical Test Theory

There are three main assumptions in the classical test theory. The first is that the error and the true scores from the same test 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 (Lord, 1980). ie $\gamma_{Te} = 0$ …… (i)

The second assumption is that the error terms have an expected mean of 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 (X = T), $^{n}\sum_{i} \frac{E}{N} = 0$ ……. (ii)

The third assumption is that the errors from parallel measurements are uncorrelated. Lord (1980) went further to posit that in the definition of parallel measurements in CTT, two measures 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.

$X \parallel X^1$ if $X_1 = X_2 = T_i + E_i$

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). 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).

$$\text{Limit } \gamma_{2} = 0$$
$$\gamma_{ij} \rightarrow 1$$

Limitations of the Classical Test Theory

The limitations of the Classical Test Theory are very prominent (Hambleton, Swaminathan and Rogers, 1991). The first limitation is that the item statistics such as item difficulty and item discrimination depend on particular examinee samples from which the test was administered. Item parameters are not invariant characteristics of the item, but take on values that depend on who tried the items. This means that characterization of an item or test is examinee (sample) dependent. Secondly, the definition of reliability in CTT is established through the concept of parallel tests. The concept of parallel tests is difficult to achieve in practice because individuals are never exactly the same on a second administration owing to factors such as forgetfulness, development of new skills or changes in motivational or anxiety levels. The third limitation of CTT has to do with the assumption that standard error of measurement is the same for all subjects and does not take into account variability in error at the different trait levels (Hambleton, Swaminathan and Rogers, 1991). The fourth limitation is the fact that CTT reflects the focus on test level information to the exclusion of item level information. CTT therefore deals with individual's total score and not their ability at the individual item level.

Item Response Theory (IRT)

Item Response Theory (IRT) provides an alternative to CTT as a basis for examining the relationship between item responses and the ability of an examinee being measured by the test or the scale (Hambleton and Swaminathan, 1985). IRT attempts to model the ability of an examinee and the probability of answering a test item correctly based on the pattern of responses to the items that constitute a test. IRT is able to estimate the parameters of an item independent of the characteristics of both test takers to which it is exposed and other items that constitute the test. Three prominent equations termed 1PL, 2PL and 3PL (parameter logistic) models are presently used to make predictions.

While there is only one parameter ascribed to the trait level of the individual, the task or item is often characterized by the three parameters. The individual trait level is often designated by theta (θ), which represents the amount of ability, trait or attribute level possessed by an individual. The three parameters associated with the item are discrimination power (a), the difficulty parameter (b), and the guessing parameter (c). In a cognitive task, the a-parameter indicates the degree to which examinees' response to an item varies with, or relates to their trait level or ability (Nenty, 2000). The b-parameter is the amount of trait inherent in an item. This represents the cognitive resistance of the item or task. In other words, this is the amount of trait under measurement just necessary to overcome the task or item. The c-parameter is the probability that a person completely lacking in the trait will overcome or answer the item correctly.

Assumptions of IRT

Warm (1978) identified four major assumptions of the Item Response Theory. The first relates to the assumption of any test theory which states that if the examinee knows the correct answer to the item, he/she will go ahead to answer it correctly. Without this assumption, there may not be any good reason for testing. The three other assumptions are local independence, unidimensionality and item response function (Item Characteristic Curve). These four assumptions are very important and should hold irrespective of the latent trait model used. In other words, a test data can only be useful for a latent trait model estimation only if these assumptions are met. Let's therefore discuss the assumptions extensively before going further. We shall also apply these assumptions to a typical physics test.

Item Local Independence

Local independence means that the probability of an examinee getting an item correct is unaffected by the answer given to other items in the test. Local independence 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. This means that 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. Ability, which influences responses to any

set of items in a test, is constant at a particular time of measurement. Therefore, the relationship 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.

The axiom of local independence states that the observed items are independent of each other given an individual's score on the latent trait (Vermunt & Magidson, 1996). Local independence means statistical independence. When items are statistically independent, each exhibits its quality and takes examinees' good display of ability in unfolding the characteristic function about them (Yen, 1993). Based on the assertion by Yen (1993) one can posit that in order to test items that would not violate the theory of local independence, the interaction between each item must not be high but should be as low as possible with respect to the logic of operations and manipulations.

Verifying the Assumptions of Local Independence

The assumption of local Independence states that the relationship between items in a test should be significantly close to zero. Onyeneho (2012) investigated the local independence of UTME pre-test Physics using Vista-Tetrachor software to determine the level of compliance of the items with the assumption of local item independence.

Table 1.1 shows the summary of tetrachoric correlations. Table 1.2 shows the frequency distribution of tetrachoric correlations for Physics subject. The percentage of correlation coefficients that are close to zero is 64.35%. Since a greater number of the correlation coefficients are close to zero, it was concluded that the UTME Pre-test Physics items are locally independent.

Table 1.1: Summary of tetrachoric correlations for Physics

Item No	**v1**	**v2**	**v3**	**v4**	**v5**	**v6**	**v7**	**v8**
v1	1							
v2	0.34	1						
v3	0.113	0.273	1					
v4	0.266	0.558	0.456	1				
v5	0.231	0.591	0.323	0.477	1			
v6	0.202	0.439	0.309	0.439	0.217	1		
v7	0.104	0.278	0.203	0.224	0.193	0.319	1	
v8	0.275	0.467	0.36	0.253	0.369	0.351	0.156	1
v9	0.281	0.471	0.22	0.276	0.476	0.375	0.175	0.501
v10	0.181	0.211	0.106	0.25	0.195	0.123	0.025	0
v11	0.345	0.441	0.365	0.379	0.421	0.144	0.18	0.311
.								
v50								

Table 1.2: Frequency Distributions of Tetrachoric Correlations for Physics Subject

Subject	Correlation Coefficient	Frequency	Percentage	Remark
Physics	Greater than 0 .500	52	4.0	
	.0.450 - 0.500	160	12.55	
	0.300 – 0.449	242	18.99	
	0.200 – 0.299	266	20.87	Close to zero
	0.0 – 0.1	554	43.48	Very close to zero

Unidimensionality

The theory of latent trait generally assumes that a set of traits underlie test performance. The examinee's ability in a set of n-dimensional latent space can be represented by a vector of ability scores as (θ_1, θ_2, θ_3….. θ_n). Item response models which assumes a single latent ability is referred to as unidimensional. Unidimensionality means that the items measure one and only one area of knowledge or ability. A set of test items testing bit of knowledge which it logically and sequentially related may be expected to be unidimensional. The condition of unidimensionally does not portend that the items must correlate positively with each other. An item may correlate negatively with others and still be unidimensional. The unidimensionality concept therefore requires that all the items on a test or ability scale must measure a single latent trait of an individual and violation of this assumption, would lead to serious misleading result (Hulin, Drasgow and Person, 1983).

JAMB Nigeria conducted a study on the 2010 UTME Mathematics test to determine possible unison or convergence of the methods in verifying or determining the asuumption of unidimensionality.

In a recent paper presented at the 2012 annual conference of IAEA in Kasakhstan by Ojerinde and Ifewulu (2012), eleven methods for testing for unidimensionality was cited. These include: The 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. Analysis was carried out using the SPSS package.

Verifying the Assumption of Unidimensionality

1. **Cronbach Alpha Test**

In the Cronbach Apha Test using the 2010 Mathematics items, a reliability alpha value of 0.943 was observed which is greater than the specified standard of 0.70

Table 1.3: Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.943	.943	49

The result of the congruence test further showed more compliance. The coefficient of congruence between group A and B yielded C_{AB} = 0.16816. This result was close to zero confirming that the 2010 UTME Mathematics items conformed with the assumptions of unidimensionality (Pine, 1977).

In the same study, the set of examinees data was separated into two gender groups and factor analysis was carried out on the two sub-groups. The first item factor loadings produced on the two sub-groups were subjected to the congruency test. The coefficient of congruence C_{AB} of the item first factor loadings between the two groups was determined using the formular:

$$\mathbf{C_{AB}} = \sqrt{\frac{\sum_{i=1}^{n} (L_{ia} - L_{ib})^2}{n}} \text{ ; as } C_{AB} \text{ then Unidimensionality} \rightarrow \infty$$

$$\mathbf{Lmt \rightarrow 0}$$

where

L_{ia} = loading of item i for group A on the first factor

L_{ib} = loading of item i for group B on the first factor

n = Number of items in the test

$C_{A,B}$ = Coefficient of Congruence between groups A & B

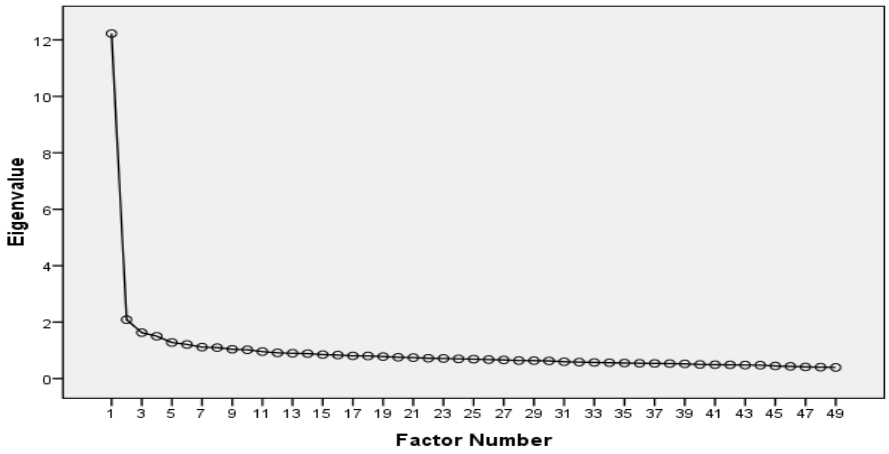
The coefficient of Congruence between A and B, therefore yielded $C_{AB} = 0.16816$.

This result was close to zero showing that the 2010 UTME mathematics items conformed to the assumption of unidimensionality.

2. Eigen Value Test

The result of the eigen value test produced the scree plot of the eigen value where the eigen value was larger compared to the second factor, and the eigen value of the remaining factors are all about the same.

Fig 1.1: Scree Plot of 2010 UTME Mathematics



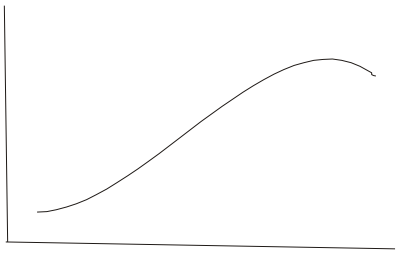
3. Factor Loading

The result of the factor loading test shows that almost all the 50 items in the 2010 UTME were loaded in the first factor (see Appendix). Five items were loaded in factor 2, while 3 items were loaded in factor three. Unidimensionality is indicated if the first factor loadings for all the items are significant and have the same sign + or – (McBride and Weiss, 1974). Furthermore, if the first eigen value is substantially greater than the next, the factor structure is deemed to have sufficiently satisfied the assumption of Unidimensionality, Orlando, Sherbouve, and Thissen (2001).

Item Response Function (IRF)

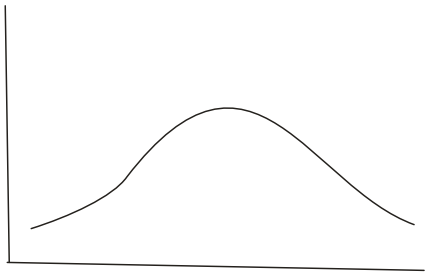
Item Response Function takes the form of the normal *ogive* (half of normal curve). The logistic frequency curve or Item Response Curve has a mean of 0 and S.D greater than 1. The normal curve has a mean of 0 and an S.D of 1. The Item Response Function is also called the Item Characteristic Curve.

Fig: 2.1



$\bar{X}$ = 0, SD > 1

Fig: 2.2

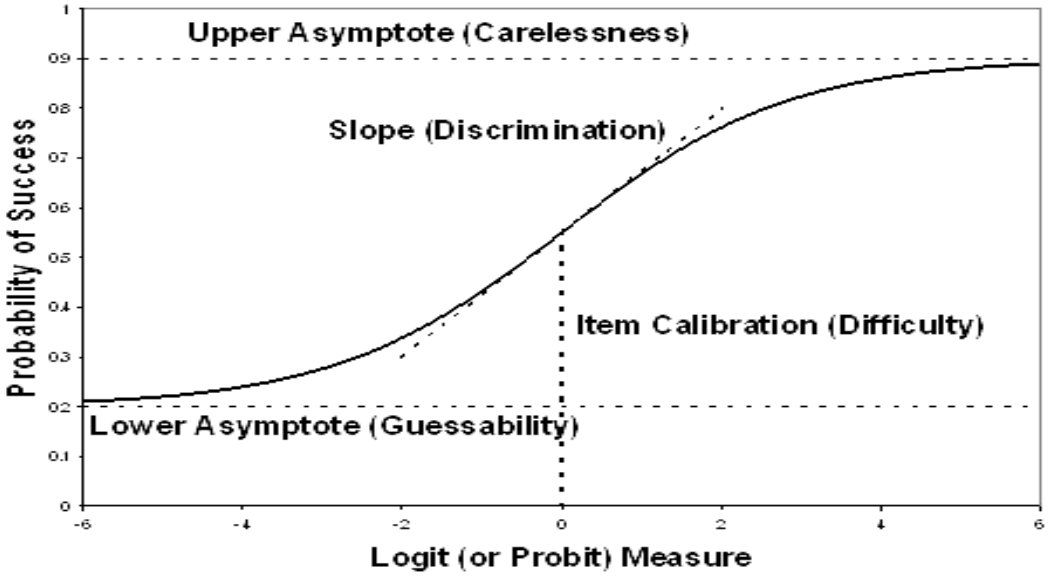


$\bar{X}$ = 0, SD = 1

When the latent trait is multi-dimensional, it is called the item characteristic function. An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test containing the item. Item characteristic curve is a non-linear regression function of item score on the latent trait measured by the test. The relationship will be non-linear as a variable is unbounded and probability is bounded.

An item characteristic curve represents the probability of a correct answer to an item expressed as a function of ability. 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 ICC is monotonic and takes the form of a normal ogive. There are three parts to it, 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 depends on the particular logistic model.

Figure 2. IRT Item Characteristic Curve



Verifying the Item Characteristic Curve Assumption

Ojerinde, Popoola and Onyeneho (2011) investigated the comparison between CTT and IRT using UTME Physics Pre-test. The ICCs of three items- 1, 11 and 22 are shown in Figures 3., 3, 4 and 5.

The values used to create the plot, i.e., the x-, y-coordinates are either a matrix or a list in which the first column or element provides the latent variable values used, and the remaining columns or elements corresponding to probabilities, information or loadings. The plotted curve is referred to as the Item Characteristic Curves (ICC) or the Item Information Curves (IIC).

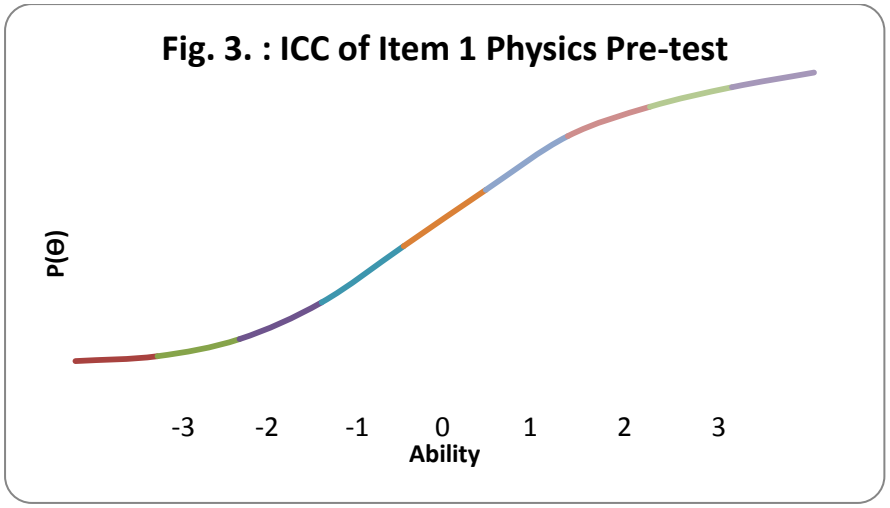


Fig. 3. : ICC of Item 1 Physics Pre-test
P(Θ)
-3
-2
-1
0
1
2
3
Ability

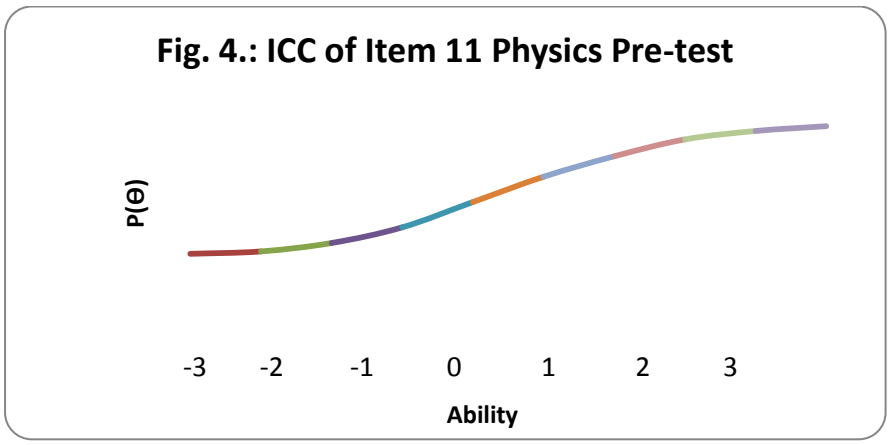


Fig. 4.: ICC of Item 11 Physics Pre-test
P(Θ)
-3
-2
-1
0
1
2
3
Ability

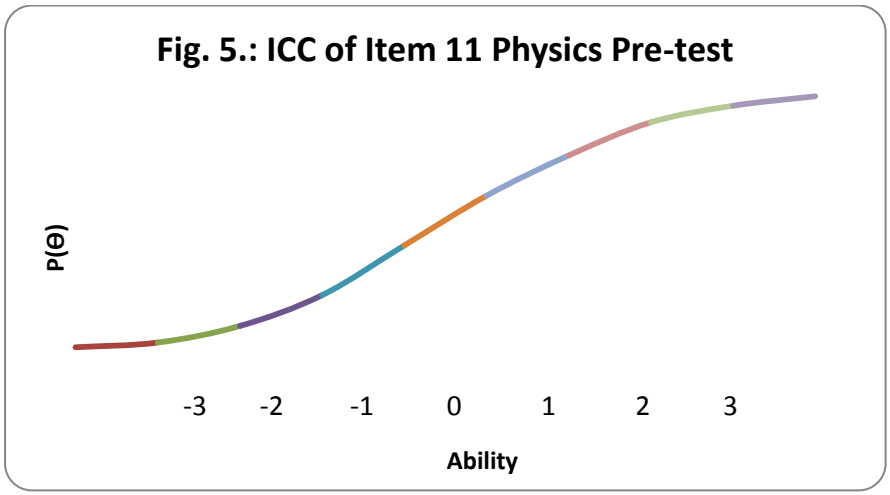


Fig. 5.: ICC of Item 11 Physics Pre-test
P(Θ)
-3
-2
-1
0
1
2
3
Ability

Comparison of CTT Item Analysis Results with that of IRT

Table 2.1: Comparison between the Item Statistics obtained from the Classical Approach and that of the Item Response Theory Model for Physics Pre-test

Item No	Analysis Using CTT			Analysis Using IRT			
	b	a	N	b	a	C	N
1	.26	.16	69	2.98	.82	.23	69
2	.26	.21	69	3.0	.84	.21	69
3	.26	.05	69	3.0	.82	.23	69
4	.63	.58	69	1.74	.80	.23	69
5	.47	.26	69	2.25	.80	.24	69
6	.42	.16	69	2.30	.82	.23	69
7	.74	.42	69	1.31	.83	.24	69
8	.32	.21	69	2.87	.82	.23	69
9	.53	.47	69	1.65	.80	.23	69
10	.79	.47	69	1.03	.76	.24	69
11	.53	.32	69	1.55	.77	.24	69
12	.74	.42	69	0.47	.80	.24	69
13	.05	.05	69	3.00	.85	.22	69
14	.37	-.05	69	2.86	.78	.25	69
15	.47	.37	69	2.53	.78	.24	69
16	.53	.26	69	1.85	.76	.24	68
17	.74	.53	69	0.87	.80	.24	68
18	.58	.21	69	1.25	.77	.25	68
19	.74	.58	69	1.30	.81	.23	68
20	.42	.32	69	2.12	.81	.23	68
21	.37	.11	69	2.74	.79	.24	68
22	.21	.21	69	3.00	.82	.23	68
23	.53	.32	69	2.11	.77	.24	68
24	.89	.63	69	0.71	.80	.24	68
25	.21	.11	69	3.00	.82	.23	68
26	.68	.53	69	1.68	.79	.24	68
27	.63	.53	69	1.96	.77	.24	68
28	.47	.32	69	1.97	.84	.23	68
29	.37	.16	69	2.74	.81	.24	68
30	.32	.00	69	2.65	.79	.25	68
31	.42	.26	69	2.69	.81	.23	68
32	.47	.37	69	2.15	.83	.22	68

33	.32	.05	69	2.66	.82	.23	67
34	.16	-.05	69	3.00	.82	.23	67
35	.11	.05	69	3.00	.83	.22	67
36	.21	-.11	69	3.00	.81	.25	66
37	.63	.42	69	1.43	.80	.24	66
38	.53	.37	69	1.64	.82	.23	66
39	.37	.16	69	2.53	.78	.25	66
40	.58	.47	69	1.69	.78	.24	66
41	.21	.16	69	3.00	.82	.23	66
42	.47	.32	69	2.23	.80	.24	66
43	.42	.32	69	2.56	.82	.23	66
44	.58	.53	69	1.74	.78	.24	66
45	.16	.05	69	3.00	.84	.22	66
46	.37	.32	69	2.56	.82	.22	66
47	.58	.32	69	1.79	.76	.24	66
48	.26	-.16	69	2.95	.79	.25	66
49	.42	.16	69	1.88	.76	.24	65
50	.37	.11	69	2.30	.77	.24	63

Table 2.2: Comparison of CTT and IRT Test Statistics for the Physics Pre-test

	Reliability	Average Information	SD	SEM	No of Items rejected because of a < .3 (IRT) or a is ($< .3$ *or* $\leq .7$) for (CTT)	No of items rejected because b is > 2.95 (IRT) or b less than .2 for (CTT)	No of Items Rejected because of keying in error (K)
CTT	.490	14.91	5.53	0.16	12	19	0
IRT	.674	3.738	a = 0.12 b= 1.0 c= 0.02	a=0.56 b=0.63 c=0.0	0	12	9

Tables 2.1 and 2.2 depict the item statistics derived from the Classical Test Theory approach and the Item Response Theory model, side by side, showed that there is a great improvement in the pre-test item statistics using the Item Response Theory compared to the Classical Test. The total number of items rejected on the basis of the discrimination index was 19 for the classical approach while only 12 were rejected using the Item Response Theory model. It can also be seen that since IRT method is sample independent, no item was rejected on the basis of item facility (item difficulty). But, because CTT is dependent on the sample used, many items (12) were rejected by the classical approach with their difficulty indexes below the bench mark of (< 0.3 or ≤ 0.7).

Again, the reliability coefficients derived using the IRT and CTT differs despite the fact that both used the Kuder-Richardson 21 formula. The reliability coefficient using IRT X-calibre programme was given as 0.674, while that of CTT was calculated as .490. The lower reliability index derived from the CTT approach is simply as a result of the rejection of more items by CTT of items with low p-values (items with low percentage correct answers) than the IRT. Worse still are items with negative discrimination values (D). That could only occur when, for some reason, the item is missed by the top test takers and answered correctly by the bottom ones.

Tables 2.1 shows that items 14, 34, 36 and 48 had negative discrimination indexes and this would have accounted for the low reliability encountered. In test development and item evaluation therefore, such items should be modified, dropped or replaced. The overall result has indicated that the IRT method used in analyzing the Pre-test data on UTME Physics was more reliable than the CTT approach.

Discussions

The results of investigations conducted on Physics Pre-test were found to be consistent with other research findings in Nigeria. An investigation was conducted by Joshua, Ubi & Abang (2011) on Item Local Independence in selection examination in Nigeria: Implications for assessment for regional integration. In the study, it was found that the UME Mathematics items for the 2000, 2001, 2002 and 2003 years were to a great extent, locally independent. A significant number of correlations were approximately zero implying that such items are not related and may not have acted as a clue to one another during the testing session.

Idowu, Eluwa & Abang (2011) carried out a comparison between CTT and IRT on Mathematics achievement test in other to determine the quality of an assessment. They found out that IRT offers a sound alternative to the classical approach. They opined that because CTT is rooted in a process of dependability rather than measurement, it does not rely on item difficulty variable for precision and calibration or on total score for indicating the measured ability (Sirotnic, 1987). Therefore, they concluded that the weaknesses of CTT have caused IRT to gain the attention of researchers since it makes allowances where CTT does not (De Ayala, 1993; Welch & Hoover,1993).

Adedoyin (2010), investigated the invariance of person parameter estimate based on classical test and item response theory. Five subsets consisting of 5 items each from 11 items that fitted the IRT models were used. The result showed that the p-values for all the raw score values was less than alpha level 0.05, which was very significant. This means that the person parameter estimate based on the 5 subsets of the 11 items were significantly different. The researcher concluded that subsets of test items developed to measure the same ability have significant influence on the estimate of such ability for the examinees. This implies that the examinees score or ability is dependent on the particular set of items administered, that is, it was test dependent.

In another study carried out by Ojerinde, Popoola & Onyeneho (2012) comparing Classical Test Theory and Item Response Theory using experience from 2011 Pre-test in the Use of English Language Paper of the Unified Tertiary Matriculation Examination (UTME), the

item statistics derived from the Classical Test Theory approach and the Item Response Theory model showed that there is a great improvement in the pre-test item statistics using the Item Response Theory over the Classical Test approach. It can also be seen that since IRT method is sample independent, no item was removed on the basis of item facility (item difficulty). But, because CTT is dependent on the sample used, many items were removed by the classical approach with their difficulty indexes below the bench mark of (> 0.3 and ≤ 0.7).

All these findings have shown that although CTT has an advantage of simplicity, the sample dependency of item and test statistics limit their usefulness and utility in psychometric analysis.

Despite the theoretical differences between IRT and CTT, there is a lack of empirical knowledge about how, and to what extent, the IRT and CTT-based item and person statistics behave differently. The findings in this study have indicated that the person and item statistics derived from the two measurement frameworks are quite comparable. The degree of invariance of item statistics across samples, usually considered as the theoretical superiority of IRT models, also appeared to be similar for the two measurement frameworks but the IRT estimation provided better reliability and functionality. The circular dependency of CTT on samples poses some theoretical difficulties in its application in some measurement situations. This is why it cannot be used effectively in test equating and computer adaptive testing (Fan, 1998). IRT has however, been used effectively in test score equating and Item Banking.

Applications of IRT to Computer Based Testing (CBT)

Estimation of item and person parameters produces more stable and precise values using IRT. Computer-based testing offers more precise traits estimation using the Item Response Theory. IRT contributes to the flexibility of CBT which offers the examiner the opportunity to administer multiple forms of a test to population or subpopulation of interest by using appropriate calibration or pertinent item characteristics such as difficulty, discrimination and other psychometric parameters which can be electronically developed for testing on demand.

The quantity of items available for test administration can also be increased by randomized and intelligent item selection both of which have added advantage of minimising answer copying in a CBT environment. As items are pseudo-randomly selected in a Linear-On-the Fly Test (LOFT), the IRT parameters are used to deliver tests to each examinee and monitor the psychometric characteristics of the test in real time. The results are then compared to psychometric targets with IRT values defined well in advance through pretesting. Sequential test are also made more achievable using IRT information. Sequential tests might be used in schools to make pass – fail decisions, in an employment context to make a decision to hire or not to hire, or in a professional certification program to determine whether an individual meets specified certification criteria. In all these decisions, IRT has proved as a valuable analysis tool.

Although some sequential tests use random item selection, the more effective tests use intelligent item selection to the extent that psychometric information on test items is used to

order items prior to item delivery. Thus, given the fixed item order, items are administered and scored one at a time. After each item is administered, a classification algorithm, such as sequential probability ratio test (Eggen, 1999, Reckage, 1983), is used to attempt to make a classification of the examinee. If a classification can be made within prescribed error tolerance, test administration is terminated for the examinee. If a high confidence classification can not be made, the next item is administered and the decision criteria again re-evaluated. This process is well enhanced with pre-calibrated tests that use a fixed order of items and allow only test length to vary. The more advanced versions of Computer Adaptive Test (CAT) allow each examinee to start their tests with different items and to receive quite different sets of test items. The most flexible and therefore efficient CATS are the fully adaptive CATS based on IRT Item Information Function which are transformations of IRT parameters. The use of information functions allows each examinee to start their test with a different item if valid prior information is available. Then based on their answers to the items, examinees are placed at a point on the IRT trait scale (Θ). This takes into account which answer the examinee gave to the item (correct/incorrect, or which rating scale alternative was selected) and the item parameter for that item. The updated score is then used to select the one unadministered item out of an entire bank that provides the most information for the examinee, which is also the item that maximally reduces the uncertainty associated with the Θ estimates. One or more termination criteria is then selected – these are typically a specified minimum value of standard error and/or some maximum number of items. If the examinee has not met one of the termination criteria, the current Θ estimate is used to select the next best item and the process continues. When a termination criteria is met, the test is ended and the final Θ estimate and its standard error are recorded for that examinee. All these are easily accomplished with the Item Response Theory parameter estimates.

Recommendations

From the foregoing, it is therefore recommended that examining bodies using multiple-choice test instruments should employ the use of both IRT and CTT statistics in test development validation processes. The invariant property of IRT model parameters makes it theoretically possible to solve some important measurement problems that have been difficult to handle within the CTT framework. Where IRT is used in carrying out such tasks, effort should be made to ensure that the assumptions of the model are tested for conformity. There cannot be full employment of the principles of IRT in solving measurement problems until the conditionalities are fulfilled. Test items have to have the property of local independence, unidimensionality and the ICCs that fit the model. The shortcomings of CTT make it imperative for the advent of IRT in determining the psychometric utility of test items and solving measurement problems. Rather than completely jettisoning CTT for IRT, it is recommended that the two approaches should form the basis of analysis. The defects of CTT can easily be compensated for by that of IRT and made to be complementary to it.

Finally, educational institutes have a role to play in advancing more interest and awareness in the application of Item Response Theory by making IRT core modules in undergraduate

programmes. Educational institutes should also sponsor graduate research and retraining of staff through local and international workshops and seminars.

Conclusion

Although CTT's major focus is on test-level information, item and test statistics are also an important part of the IRT model. At the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item if the items are dichotomously scored.

On the other hand, the IRT model focuses on both item and person statistics such as the item parameters, item characteristic function, test information characteristics and test information functions. These features portray IRT as a better option in giving adequate information concerning the behaviour of an item as well as the test takers.

References

Adedoyin, C. (2010). Investigating the Invariance of Persons Parameter Estimates based on Classical Test and Item Response Theories 2010. An international journal on education science 2 (2);107-113

Adema. J. J. (1990) The Construction of Customized Two-Stage Tests Journal of Educational Measurement *Fall 1990, Vol. 27, No. 3, pp. 241-253University of Twente*

Birnbaum, A. (1968) "Some Latent Trait Models and their Uses in Inferring an Examinee's ability". In F. M. Lord, and Mr. Norvick, statistical theories of mental test scores. Reading MA: Addison – Wesley

Fan, X. (1998) Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement

Hambleton , R.K., Swaminathan , H.,& Rogers, H. J. (1991) Fundamentals of Item Response Theory. Newbury Park, CA:Sage Publications

Hambleton, R. K (1989) "Principles and Selected Applications of Item Response Theory". University of Massachusettes at Amherst

Hambleton, R.K. and Swaminathan, H. (1985)Item Response Theory .Principles and Application.

Hambleton, R.K., Swaminathan, H.,& Rogers, H.Jl(1991).Fundermentals of Item Response Theory. New buryPark .,CA ;Sage Press

Hulin, C. L., Lissac, R. I., and Drasgow, F. (1983) Joint Maximum Likelihood Estimate# Fish girl on scribd www.scribd.com/doc/19444534/Joint Maximum Likelihood...

Hulin,C. L., Drasgow, F., &Parsons, C. K. (1983) Item Response Theory; Application to psychology www. Brain bench .com

Idowu, E. O., Eluwa, A. N., & Abang, B. K. (2011) "Evaluation of Mathematics Achievement Test: A Comparison between Classical Test Theory (CTT) and Item Response Theory (IRT). Department of Educational Foundation, Guidiance and Counselling. University of Calabar, Calabar, Cross River State, Nigeria

Lord, F. M. (1980)"Application of Item response Theory to practical testing problems". Hillsdale, NJ:Erlbaum

Lord, F.M. (1980) "Application of Item Response Theory to Practical Testing Problems". Hillsdale , NJ: Lawrence Erlbaum.

Lord, F.M.(1953) An application of confidence interval and maximum likelihood to the estimation of an examinee's ability .

Lumbersden, J.(1957) "A Factorial Approach to Unidimensionality". *Australian Journal of psychology 9,105-111*

MCBride, and Weiss, D.J. (1974), p.37 Calibration of an item pool for adaptive

Ojerinde D. and Ifewulu C.B (2012) Item unidimensionality using 2010 UTME Mathematics a paper presented at IAEA conference in Astana Kasakstan

Ojerinde, D., and 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

Ojerinde, D., Popoola, K., Ojo, F., and Onyeneho, O. P. (2012)Introduction to Item Response Theory: Parameter models, estimation and application. Goshen Print media Ltd

Onyeneho, O. P. (2012) Ensuring Item local independence in the Unified Tertiary Matriculation In the Application of IRT for Regional Development in Africa. An unpublished paper

Orlando, M., Sherbourve, C. D. & Thissen, D (2001). Summed-score linking using Item Response Theory: Application to depression measured. Psychological Assessment, 12(3), 354 – 359.

Pine, S.M. (1977). Applications of Item Response theory to the problem of test bias. In D.J. Weiss (Ed.), Applications of computerised adaptive testing (Research Report 77-1). Minneapolis: University of Minnesota, Psychometrics Methods Program, Department of Psychology.

Reckage, M.D. (1985) The difficulty of test items that measure more than one ability. Applied Psychological Measurement. 5, 11-19.

Vermunt, J. K, & Magidson, J. (1996) "Graphical Displays for Identified and Unidentified Latent Class Cluster and Factor Models".

Warm, T. A. (1978) Primer of Item Response Theory. U.S Coast Guard Institute, Oklahoma city Oklahoma

Wikipedia The Free Enclopedia, 2011

Yen, T.Y. (1993) "A comparison of three statistical procedure to identify clusters of items with local dependency". Huynh University of Carolina

Appendix

Factor Matrix of 2010 UTME Mathematics

	Factor								
	1	2	3	4	4	5	6	7	8	9
R2	.444	-310								
R3	.510	-330								
R4	.622									
R5	.524									
R6										
R7	.590									
R8	.506				.328					
R9	.563									
R10	.493									
R11	.435									
R12	.618									
R13	.507									
R14	.437									
R15	.456									
R16	.546									
R17	.599									
R18	.638									
R19	.503									

R20	.462							
R21	.606							
R22	.517							
R23	.623							
R24	.475							
R25	.593							
R26	.480							
R27	.306	.401						
R28	.427							
R29	.371							
R30								
R31	.496							
R32	.416	.383						
R33	.372	.392						
R34			.333					
R35	.516							
R36	.565							
R37								
R38	.487							
R39	.465							
R40	.398							
R41	.349		334					
R42	.521							
R43	.587							
						.322		
R44	.461							
R45	.588							
R46	.457							
R47	.568							
R48	.463							
R49	.438							

R50	.363							

Extraction Method: principal Axis Factoring.
a.9 factors extracted. 11 iterations required.